IEEE Access

Received 4 November 2024, accepted 2 December 2024, date of publication 5 December 2024, date of current version 17 December 2024.

Digital Object Identifier [illegible]

SURVEY

Arabic Text Formality Modification: A Review and Future Research Directions

SHADI I. ABUDALFA, **FAHAD J. ABDU**, AND **MAAD M. ALOWAIFEER**

SDAIA-KFUPM Joint Research Center for Artificial Intelligence (JRC-AI), King Fahd University of Petroleum & Minerals, Dhahran, Eastern Province 31261, Saudi Arabia

Corresponding author: Shadi I. Abudalfa (shadi.abudalfa@kfupm.edu.sa)

This work was supported by the Saudi Data & AI Authority (SDAIA) and KFUPM through the SDAIA-KFUPM Joint Research Center for Artificial Intelligence under Grant JRC-AI-UCG-05.

ABSTRACT Formality transfer seeks to adjust text formality without altering its core meaning, which carries substantial implications across diverse domains like machine translation, dialogue systems, and social media content creation. This study provides an extensive overview of formality transfer specifically within Arabic text, an emerging domain within natural language processing. Particularly, we carried out a comprehensive review of literature on text formality transfer, focusing on studies published between July 2010 and April 2024. Our focus lies in treating formality transfer in Arabic as akin to a machine translation task, presenting synthesized insights. Despite advancements in formality transfer for English and other languages, Arabic's distinct linguistic features present unique challenges and opportunities. Our investigation uncovers several research gaps necessitating future exploration, emphasizing persistent limitations. Moreover, we delve into text formality transfer as a promising avenue for forthcoming research initiatives in the realm of Arabic text processing.

INDEX TERMS Style transfer, formality, Arabic text, dialect, machine translation.

I. INTRODUCTION

In the era of digital communication, the manipulation of textual content for diverse objectives has seen a notable rise [1]. Of specific concern is the modulation of styles [2] embedded within text. Text style transfer [3] encompasses the conversion of textual content from one stylistic format to another with the objective of preserving its semantic essence while adjusting its expressive elements. This procedure involves adjustments to linguistic aspects like word choice, sentence structure, and mood to achieve the intended style with minimal deviation in the text. In this study, we concentrate on choosing formality as the attribute for text style transfer.

The process of formality transfer entails modifying the stylistic rendition of a sentence, transitioning it from an informal tone to a formal one, while preserving its inherent meaning. This procedure bears substantial implications for various applications, including machine translation (MT) [4], dialogue systems [5], [6], [7] through online social media [8], and review systems, emphasizing the critical importance of comprehending and producing text with suitable formality levels.

There is an abundance of Arabic dialects (ADs) [9] in the Arabic world. Nevertheless, a formal dialect exist referred to Modern Standard Arabic (MSA) [10]. Whereas, Arabic dialect (AD) is an informal form of MSA language spoken across all Arabic countries. MSA serves as the academic discourse and communicative variation of the Arabic language, acknowledged by the United Nations. The high prevalence of ADs in Arabic-speaking countries stimulates the need for research focused on creating automated systems capable of implementing text formality transfer.

The primary objective of Arabic text formality transfer entails transforming text from one of ADs to the MSA. The concept of Arabic text formality transfer is depicted in a schematic illustration in Fig. 1. The diagram additionally highlights the key stages utilized in creating a system for formality transfer in Arabic text. The initial critical phase involves gathering Arabic sentences from various sources,

The associate editor coordinating the review of this manuscript and approving it for publication was Maria Chiara Caschera.

FIGURE 1. General representation of the Arabic text formality transfer.

including social media platforms. This phase also involves annotating data to create a parallel dataset comprising ADs and their corresponding MSA. The second step involves preprocessing the raw data, which includes tokenization and cleaning up special characters by removing unnecessary elements like URLs and HTML tags. The model or technique is then designed to convert informal text into formal MSA. Lastly, the proposed technique's performance is assessed to finalize the development life cycle.

Neural machine translation (NMT) has achieved considerable progress in translating languages with ample resources, such as English, thanks to the availability of large-scale corpora that supply extensive training data for NMT models. However, languages with fewer resources, including Arabic and its various dialects, encounter significant challenges in MT due to the limited amount of text data necessary for developing effective translation models [11]. Over the past two decades, the field of MT has experienced remarkable advancements, resulting in improved translation quality, though it has not yet reached the level of human translation. Despite this progress, the effectiveness of MT systems is further hampered when translating ADs, and research into MT between ADs and MSA has been relatively sparse.

Within this manuscript, we delve into the realm of Arabic text formality transfer, analyzing current approaches, hurdles, and prospective pathways. Through a comprehensive analysis of used techniques and resources, we aim to provide insights and frameworks that advance the field of Arabic natural language processing (NLP).

This study does not consider the transformation in the reverse direction: from MSA to AD [12]. We also exclude in this study the transliteration of Arabic dialects written in Latin script (known as Arabizi) into MSA [13]. We excluded as well code-switching style transfer [14].

As far as we are aware, no prior systematic literature review has been conducted in this research direction, which drives the motivation for our current study. Our study enriches the broader conversation regarding linguistic diversity, intercultural dialogue, and the manipulation of digital text, emphasizing the role of formality adjustment in improving the richness and variety of Arabic textual materials. The contributions of this research are based on:

- Covering the latest research findings from 2010 to 2024 and a substantial volume of papers derived from a well-defined literature review procedure.
- Presenting a taxonomy framework for organizing the body of literature in AD to MSA machine translation, along with categorizing various methods employed for Arabic text formality transfer.
- Classifying 145 chosen articles based on the suggested taxonomy.
- Summarizing the latest formality transfer models, emphasizing machine learning techniques.
- Proposing new avenues for exploration in future research aimed at enhancing approaches for Arabic text formality transfer.

The structure of the paper unfolds as follows: Section II discusses earlier studies on literature reviews pertaining to formality transfer and associated research trends. In Section III, we describe the methodology employed in conducting this survey. Section IV presents our findings with this study. Additionally, Section V furnishes discussions and suggests avenues for future research. Finally, in Section VI, we draw conclusions.

II. RELATED WORKS

Arabic text formality transfer can be integrated with other research avenues to advance larger systems for Arabic NLP. The following subsections discuss related works that align with our research focus.

A. RELATED REVIEWS

Some survey studies can be associated with developing systems for transferring Arabic formality from various perspectives. In this subsection, we present several reviews that are relevant to our research direction from various perspectives.

For example, Elnagar et al. [15] encompass computational approaches to dialectal Arabic detection. The findings indicate an uneven distribution of research efforts between speech and text and among the dialects, with a bias towards text over speech, regional varieties over individual dialects, and a preference for Egyptian dialects. The review also highlights significant research gaps related to the complete neglect of some Arabic dialects, a lack of resources for city dialects, and a deficiency in deep machine learning experimentation. This study diverges from our work as it concentrates exclusively on detecting dialects. Nonetheless, dialect detection can be utilized to create an automated formality system that identifies the specific Arabic dialect and subsequently translates the text into MSA. Additionally, our study is concentrated solely on the text modality.

Harrat et al. [16] center on MT within the realm of Arabic dialects. It offers an overview of contemporary research in this field, presenting a comprehensive account of each study's methodology and highlighting its key contributions. It is important to note that this study is quite dated, having been submitted to the journal on August 23, 2017. Therefore, our current research is highly significant as it demonstrates the advancements in this field following the recent renaissance of large language models (LLMs).

Alqudsi et al. [17] aim to highlight various MT techniques found in the literature to inspire further research in this area. They examine specific linguistic features of the Arabic MT and addressing potential challenges. Additionally, they provide a detailed overview of the primary methods used for translating Arabic into English, evaluating their respective strengths and weaknesses. This study has a limitation in that it does not provide detailed information on translating Arabic dialects. Moreover, it is an older study, having been published online on July 24, 2012. Similarly, Sufiya et al. [18] explore a range of translation models, such as Convolutional Neural Networks (CNNs), LSTM, NMT, BERT, and novel hybrid structures like the Transformer-CNN. The study highlights the advantages and drawbacks of each model, revealing their capacities to tackle translation issues.

In addition, Elsherif and Soomro [19] aim to examine the approaches, challenges, and proposed solutions in Arabic MT from approximately seven years ago. In the same context, Almaaytah and Alzobidy [20] identified nineteen challenges from 56 research papers. Their study delves into the four most significant issues and explores the solutions proposed by other researchers. The primary challenges include word sense disambiguation, Arabic named-entities, intricate morphology, and limited resources.

Moreover, Zakraoui et al. [21] conduct a thorough examination and contrast of various NMT methodologies, primarily focusing on Arabic-English MT endeavors. The methodologies under discussion tackle diverse linguistic and technical hurdles, showcasing significant advancements over conventional approaches. The objective of their inquiry is to ensure that researchers are informed about the latest research advancements until 2021, providing them with vital materials to improve Arabic MT. In a similar context, Ameur et al. [22] furnish an overview of significant research endeavors in Arabic MT, highlighting key studies and available resources for developing and evaluating Arabic MT systems.

TABLE 1. Comparison with related reviews.

Reference	Year	Machine translation	Dialect identification	Dataset(s)	Evaluation metrics	Formality transfer	Comparisons	Source language	Target language
[17]	2016	✓	×	×	✓	✓	✓	MSA	Eng
[19]	2017	✓	✓	×	×	×	✓	MSA	Eng
[16]	2019	✓	×	×	×	×	✓	ADs	Eng
[22]	2020	✓	×	✓	×		✓	MSA	Eng
[2]	2020	×	✓	×	×	×	×	All	All
[23]	2020	✓	×	×	×	✓	×	ADs	MSA
[15]	2021	×	✓	×	✓	×	✓	ADs	All
[21]	2021	✓	×	✓	✓	✓	✓	MSA	All
[20]	2022	✓	✓	×	×	×	✓	MSA	Eng
[18]	2024	✓	×	×	✓	×	×	MSA	Eng
Ours	2024	✓	×	✓	✓	✓	✓	All	All

Zampieri et al. [2] explore pertinent applications like language and dialect identification, as well as MT, particularly concerning closely related languages, language variations, and dialects. Their study delves into computational techniques for handling languages with similarities but does not specifically concentrate on Arabic dialects. Within this context, Alnassor et al. [23] focus is on Arabic and its dialects, considered low-resource languages, particularly in the context of converting non-standard text through normalization and translation techniques.

According to the results of this subsection, it is evident that our survey diverges from the related literature in several respects (refer to Table 1). Hence, according to our understanding, our study marks the inaugural systematic literature review concentrating on the transfer of formality in Arabic text. This study aims to shed light on contemporary methodologies for enhancing the efficacy of conversational systems in informal Arabic discourse, encouraging further investigation among scholars.

B. RELATED TASKS

Text formality transfer is generally associated with other tasks as it falls under the broader category of text style transfer. One closely related task is MT, which involves converting text from one language to another. In the following subsections, we explore how various NLP tasks relate to this study, which centers on formality transfer in Arabic.

1) MACHINE TRANSLATION OF ARABIC AND NON-ARABIC TEXTS

Many studies have been conducted over the years translating between Arabic text and other languages [24], [25], [26], Typically, MT relies on parallel datasets [27]. The studies cut across the classical methods (i.e. rule-based and statistical based techniques) and the recent NMT techniques (i.e., sequence to sequence models). Of course, the most recent LLMs generative models have now been employed in this domain for achieving exceptional translation quality. In this subsection, we discuss several studies on Arabic to other language translations to illustrate the principles of MT. The chosen studies in this subsection mainly present Arabic-English translation [28]. Therefore, they are not regarded as *Arabic* formality transfer. We highlight the distinction between Arabic/non-Arabic translation and ADs/MSA translation as elucidated in the subsequent sections of this paper.

For instance, Berrichi and Mazroui [29] developed an English to Arabic NMT model to mitigate the effects of limited vocabulary and long sentences constraint encountered in translating languages with contrasting structures (i.e., English and Arabic). Similarly, Aljohany et al. [30] designed a two-way NMT-based model between Arabic and English texts. The NMT system is based on LSTM encoder-decoder model that employed attention network. Experiments confirmed that the proposed approach led to increased in translation quality, as evidenced by a reduction in translation loss.

Sajjad et al. [31] developed an evaluation suite for translation between Arabic dialects and English (by using some public datasets such as QAraC [32]), encompassing a diverse array of dialects, spanning multiple genres, and varying levels of dialectal characteristics. The suite is the first of its kind, bringing together a wide range of dialects, spanning multiple domains, and exhibiting varying degrees of dialectal characteristics. This research group employed an industrial-scale MSA to English systems to develop robust baselines through fine-tuning, back-translation, and data augmentation techniques.

Additionally, Sidiya et al. [18] developed two distinct models for translating English to Arabic: one utilizing LSTM and the other employing BERT. An extensive performance evaluation of these models is conducted, followed by a detailed examination of the outcomes. The comparative analysis yields valuable insights into the current state of Arabic-English translation models, offering guidance for future studies to enhance model accuracy.

Hamed et al. [33] proposed a NMT system with sole objective of improving the translation quality of the Holy Quran Arabic text to Italian language. The developed model employed two sequence to sequence-based deep learning algorithms including LSTM and GRU with attention mechanism. Their experimental evaluations demonstrate that LSTM standout with an average BLEU and ROUGE performance of 0.18 and 0.17, respectively. However, their method is evaluated using small training samples, thus extending the experiments using huge amount of data needs to be conducted to ascertain the effectiveness of the developed model.

Bensalah et al. [34] proposed a NMT system between English and Arabic text that employed their novel word embedding scheme generated by combining word2vec (i.e., CBOW and SG) and FastText embeddings models. The developed embedding method was used to train CNNs and various versions of RNNs to achieve the MT. They examined the efficacy of the developed approach using UN dataset and compared it with word2vec and FastText standalone embedding models.

Alrajeh [35] presents a NMT system to translate between Arabic and English using different parallel corpora and compared them with conventional approaches such as statistical machine translation (SMT) model (i.e., phrase-based systems). The author also explored the significance of specific preprocessing steps tailored to the Arabic script. After extensive experiments, the author has established that indeed preprocessing steps help in improving the translation quality of both NMT and the traditional methods. The findings indicate that NMT surpassed the phrase-based systems in all settings. Their study also revealed that training a neural network on a limited dataset produced results competitive with those achieved by conventional systems.

Furthermore, Oudah et al. [36] performed comprehensive comparison between SMT and NMT systems for translation between Arabic and English using training samples preprocessed by different suitable tokenization approaches. Moreover, they examined various data and vocabulary sizes and assessed their impact on both methodologies. After experimental evaluations, results obtained with their work indicate that selecting the optimal tokenization scheme depends significantly on the model type and dataset size. Specifically, performance achieved from the developed method surpassed those documented in previous studies when evaluated on in-domain tests. In addition, the authors demonstrate that significant enhancements can be achieved through a system selection strategy for amalgamating the outputs from both neural and statistical MT.

Bensalah et al. [37] developed a sequence-to-sequence (Seq2Seq) deep learning algorithm to translate between Arabic and English languages. In particular, the authors conducted extensive comparison of four different encoder-decoder based model architectures including LSTM, BiLSTM, GRU and BiGRU integrated with attention mechanism. They also examined the influence of various preprocessing approaches on the developed MT system. The experimental findings revealed that the combination of Bi-GRU as an encoder, BiLSTM as a decoder, and coupled with attention network achieved the most optimal performance.

Recently, models based on transformers have demonstrated notable efficacy in language understanding and have achieved leading-edge performance across various NLP tasks, particularly MT. In this respect, Bensalah et al. [38] present NMT-based model to enhance the translating quality achieved on the Arabic-English-based translation systems. The developed approach is based on CNN and transformer architectures. The authors proposed novel preprocessing schemes that are based on FARASA and Arabert. Results obtained using UN Arabic-English corpus indicate that their developed model achieved better translation quality surpassing the current advanced Arabic based NMT methods. In the same context, Benzalah et al. [39] developed a hybrid CNN and RNN attention-based algorithm to translate between Arabic and English text achieving outstanding translation quality.

Moreover, some of related works translate mixed Arabic dialects (i.e., combination of more than one dialect in the same text) into MSA. Nagoudi et al. [40] translate MSA mixed with Egyptian Arabic dialect to English using sequence to sequence transformers (S2ST) trained from scratch and pre-trained large language model. They use Open Parallel Corpus (OPUS) [41] for conducting the experiment work.

MT task can also be achieved by exploiting or adapting some of the pre-trained LLMs such as BERT, T5 and the recently introduced Arabic-based pre-trained transformer models like AraBert and AraT5. In the same context, Nagoudi et al. [42] developed a novel evaluation benchmark for Arabic MT.

In spite of the success recorded by the NMT techniques in translating between Arabic and English text and vice versa, attention has now turned to the most recent LLMs paradigms. Over the last few years different researchers have started deploying recent LLMs like ChatGPT to ascertain their performance in various natural language generation (NLG) tasks such as MT. Specifically, the performance of these LLMs is being investigated in respect to under-resourced languages such as Arabic. For example, Al-Khalifa et al. [43] examined how the most advanced pre-trained language models (PLMs) handle MT tasks from English to Arabic. Evaluations based on various performance metrics illustrate that GPT-4 and GPT-3.5-turbo exhibit better proficiency.

Alkhedor et al. [44] present the initial research endeavor investigating translation between Syrian Arabic and Turkish. They assess translation accuracy into Turkish from different Arabic dialects and contrast these findings with translations from Syrian Arabic. To establish their MADAR-Turk dataset, they translate the aforementioned 2,000 sentences from the Damascus dialect of Syrian Arabic into Turkish, aided by two native Syrian Arabic speakers proficient in Turkish.

Although LLMs like ChatGPT and Bard are believed to possess multilingual proficiency after being fine-tuned on instructional data, their level of linguistic has not been adequately investigated. LLMs are usually evaluated with zero-shot (ZS) and few-shot (FS) [45] generation strategies. To explore this constraint, Kadaoui et al. [46] provide an extensive assessment of LLMs in MT tasks. Specifically, they assess ChatGPT, GPT-4, and Bard by translating between 10 different Arabic dialects and English. Similarly, Khondaker et al. [47] performed a comprehensive performance evaluation of LLMs with primary goal of assessing the capabilities of ChatGPT's on different NLP tasks including Arabic language and its associated dialects. The findings of their study suggest that while Chat-GPT demonstrates impressive results in English, it consistently falls behind smaller models that have been fine-tuned for Arabic (e.g., MARBERTv2 and AraT5).

Moreover, Mosleh et al. [48] investigate the functionalities of various LLMs models including GPT-3.5, GPT-4, and BLOOM for adaptive machine translation in real-time via in-context learning. Extensive experiments were conducted across five different language combinations (i.e., English-to-Arabic, Chinese, French, Kinyarwanda and to Spanish). Their findings indicate that the translation performance achieved through FS based LLMs (GPT-3.5) can exceed that of robust encoder-decoder-based MT models, particularly for English-to-French and English-to-Spanish. While, low performance was realized for English-to-Arabic.

Alkhawaja [49] assesses the suitability of ChatGPT for Arabic-English MT task. The primary aim of the developed approach is to examine the quality of ChatGPT's translations and to compare its performance with dedicated commercial MT systems like Google translate. To achieve this goal, a standardized dataset consisting of 1000 English sentences and their respective Arabic translations was utilized to assess the translation results generated by both MT systems, as well as a human translation reference. Their findings suggest a slight edge of ChatGPT over Google translate in producing high-quality translations. Nevertheless, despite these encouraging translations quality, it is important to recognize that even the most advanced MT technology, including ChatGPT, currently cannot achieve the level of proficiency exhibited by human translation. Table 2 presents a summary of MT approaches used with Arabic and other languages.

Fig. 2 displays the publication trends between 2018 and 2024. The data reveals a steady rise in publication numbers over time. This indicates interest in the field, particularly as it intensified in the latter part of the period under review.

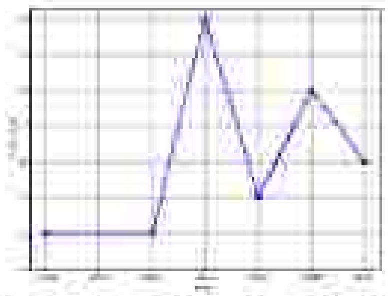

FIGURE 2. Publication distribution by year for arabic and non-arabic machine translation.

2) MACHINE TRANSLATION FOR NON-ARABIC LANGUAGES

Numerous studies have been conducted on utilizing machine translation across various languages. In this subsection, we discuss several recent studies that address machine translation involving languages other than Arabic. Recently, Lyu et al. [50] evaluated different MT tasks such as long document translation, stylized MT, and interactive MT using

TABLE 1. Summary of MT approaches used with Arabic and other languages.

Reference	Year	Models	Source	Target	Metric	Dataset	Result
[illegible]	2018	Marian NMT system	Arabic	English	BLEU	Tatoeba, MT and UN	[illegible]
[illegible]	2020	Transformer-based [illegible]	AD and MSA	English	BLEU	APT, MDPC, MADAR, Qatari and Bible	MSA-English: [illegible] All dialects to English: [illegible]
[illegible]	2022	Transformers and Pre-trained Sequence-to-sequence Language models (LMs)	AD and MSA	English	BLEU	[illegible]	[illegible]
[illegible]	2021	mT5, AraT5, [illegible]	Arabic/English, French, German, Russian	English/Arabic	BLEU	ADPT, MDPC, MADAR, Qatari, IWSLT, UN	[illegible]
[illegible]	2023	Encoder-decoder	English/Arabic	Arabic/English	BLEU	UN documents corpus	English to Arabic: [illegible] Arabic to English: [illegible]
[illegible]	2023	LSTM and GRU with attention mechanism	Holy Quran (CA)	Arabic	BLEU and ROUGE	Holy Quran text	[illegible]
[illegible]	2023	CNNs and RNNs	Arabic	English	BLEU	UN dataset	[illegible]
[illegible]	2023	LSTM, BiLSTM, GRU, and BiGRU with attention mechanism	Arabic	English	BLEU	UN dataset	[illegible]
[illegible]	2023	CNN and transformers	Arabic	English	BLEU	UN Dataset	[illegible]
[illegible]	2023	LSTM with attention mechanism	Arabic/English	English/Arabic	Accuracy and loss	Arabic-English Parallel Corpus (AEPC)	[illegible] (Arabic to English) [illegible] (English to Arabic)
[illegible]	2023	Hybrid CNN-LSTM with attention network	Arabic	English	BLEU, TER, and WER scores	Private own Arabic-English dataset	[illegible]
[illegible]	2023	GPT-3.5, GPT-4	English	Arabic	BLEU	United Nations version 1 (UN)	[illegible]
[illegible]	2023	ChatGPT GPT-3.5, GPT-4 and BLOOM	English	Arabic, Chinese, French, Spanish, Kinyarwanda	spBLEU, ChrF++, TER, and COMET	[illegible]	[illegible]
[illegible]	2023	Google Translate	Arabic	English	[illegible] Error, Semantic Error, Morphological Error and BLEU	[illegible]	[illegible]
[illegible]	2023	Bard, GPT-3.5 and GPT-4	ADs, MSA, CA	English	BLEU, COMET	[illegible]	[illegible]
[illegible]	2023	ChatGPT, GPT-4	MSA, ADs	English	BLEU	[illegible]	[illegible]
[illegible]	2024	Google Translate, [illegible] GPT-3.5-turbo and GPT-4	English	Arabic	ChrF, BERTScore, COMET	[illegible]	[illegible]
[illegible]	2024	LSTM and BERT	Arabic	English	Accuracy, BLEU, and Loss	Public Dataset	[illegible]
[illegible]	2024	ChatGPT, Google Translate	Arabic	English	BLEU	[illegible]	[illegible]

LLMs. Furthermore, they tackle privacy concerns in MT using LLMs and suggest fundamental privacy-preserving techniques to alleviate potential risks.

Additionally, Zhang et al. [51] explore the abilities of LLMs in performing MT tasks based on QLoRA [52] to translate between English and French. They employed three important techniques including prompting, in context learning and fine-tuning to accomplish the evaluation. Their study shows that the effectiveness of LLMs for MT is not always consistent. LLaMA-2 surpasses its competitors. However, models that depend completely on FS learning often underperform compared to models that are trained from scratch. The process of fine-tuning consistently yields performance improvements, particularly for models encountering difficulties with FS learning paradigms and document-level translation tasks.

Similarly, Jiao et al. [55] evaluated how well ChatGPT performs MT tasks compared to popular commercial services like Google Translate. They looked at how these systems handle different aspects of translation, including following user instructions, translation between many languages (multilingual), and adapting to different text styles. After extensive experiments they were able to establish that ChatGPT performed competitively with Google, DeepL and Tencent translation translators on high resource languages but struggle on low resource languages based on BLEU metric. The authors investigate also an intriguing approach known as pivot prompting for distant languages, wherein ChatGPT translates the candidate sentence into a high-resource pivot language prior to the conversion to target language, thereby notably enhancing translation performance. In respect to translation robustness, ChatGPT does not match the performance of commercial systems on biomedical abstracts or Reddit comments. However, it demonstrates good performance on language translation.

Moreover, Hendy et al. [56] present an extensive assessment of GPT models for MT, addressing different facets including the performance of various GPT models compared to cutting-edge research and popular commercial services, the impact of prompting techniques, resilience to changes in domain, and document level translation. Their experiments involve translation tasks in 18 diverse directions, covering both well-resourced and limited-resource languages. This assessment extends beyond English translation alone, evaluating the performance of three GPT models: ChatGPT and two different versions of GPT-3.5. The achieved performance indicates that GPT models attain translation quality that competes effectively, particularly for high resource languages. However, they exhibit limited capabilities for low resource languages. The authors also demonstrate that integrating GPT models with other translation algorithms has the potential to further improve translation accuracy.

Most of the text style transfer studies in the literature concentrate on the well-resourced languages such as English abandoning limited-resource languages such as Chinese. To address this disparity, Tao et al. [57] developed a Chinese article style transfer model, harnessing the potential of LLMs. This model integrates a Text Style Definition (TSD) module designed to thoroughly analyze textual attributes in articles, facilitating the seamless transfer of Chinese article-style by LLMs. The TSD module leverages various machine learning techniques in examining the article style. Extensive experiments confirm that the developed approach surpasses previous one in terms of transfer accuracy and maintaining content integrity.

In the same context, Dare et al. [58] explored unsupervised translation system to transform between Mandarin Chinese and Cantonese. A comprehensive comparison was conducted across various model structures, tokenization methods, and embedding techniques at a large scale. Their most successful model attained a BLEU score of 25.1 in translating Mandarin to Cantonese.

3) OTHER ATTRIBUTES FOR TEXT STYLE TRANSFER

Our research focuses on examining the techniques employed for formality transfer. Nonetheless, various methods used in related sub-tasks like politeness transfer and emotion modification can also be applied to formality transfer. For example, the impressive ability of diffusion probabilistic-based models in generating high-quality images with a high degree of control has inspired researchers to investigate the potential of applying similar control mechanisms to the domain of text generation. Prior investigations into diffusion-based models have revealed their capability to be trained without the pre-trained weights.

In this regard, Lyu et al. [59] performed style transfer by adapting diffusion-based algorithm and training it from scratch (i.e., without pre-trained weights) with very small training samples. They evaluated their approach by using StylePTB dataset (i.e., a benchmark for fine-grained text style transfers). Additionally, their developed approach, trained on a small sample of StylePTB, achieves superior performance compared to prior methods that depend on pre-trained weights, embedding layers, and external linguistic analyzers.

III. RESEARCH DESIGN

As previously stated, this study seeks to enhance our comprehension of the existing advancements in automated support for Arabic text formality transfer. It explores the literature, examines available datasets, and reviews the techniques employed in text formality transfer. Additionally, it aims to understand how researchers evaluate the quality of formality transfer.

It is important to point out that our focus in this work revolves around elucidating the challenges, datasets, and existing methodologies pertaining to the formality transfer of Arabic text. Additionally, we investigate different methodologies and present empirical results to contribute to the advancement of Arabic natural language processing research. Fig. 3 shows the main themes addressed in this study.

We formulated three Research Questions (RQs) to align with our study objectives. Table 1 presents the RQs

FIGURE 3. A framework to review studies concerning the transformation of formality in Arabic text.

TABLE 3. Research questions and the reasoning behind them.

#	Research question	Reasoning
RQ1	[illegible]	[illegible]
RQ2	[illegible]	[illegible]
RQ3	[illegible]	[illegible]

considered in our study, along with an explanation of the rationale behind each one.

In this section, we describe the review protocol steps implemented in our study. In Subsection III-A, we outline our approach to searching for relevant studies. Subsection III-B details the criteria used to systematically identify and choose research articles on text formality transfer, as well as the quality evaluation method applied to assess the selected studies. Lastly, in Subsection III-C, we describe our process for data extraction and synthesis to address the posed RQs.

A. SEARCH STRATEGY

We employed a two-stage hybrid search strategy that combined database searches with the snowballing method. Initially, we conducted a search within databases, followed by a snowballing search. A database search involves creating and executing a search query within digital libraries. A snowballing search involves examining the references (backward) and citations (forward) of chosen studies to identify additional relevant research. According to Mourão et al. [60], an effective approach for locating evidence in systematic reviews is to utilize a hybrid search strategy that combines a representative digital library with the snowballing method.

TABLE 4. Query strings used to identify primary studies.

#	Query string
1	"Arabic dialect" "machine translation"
2	text "formality transfer"
3	"Arabic text" "formality transfer"

To conduct the database search, we pinpointed appropriate keywords. By examining our RQs, we identified key concepts and terms linked to our primary topic, which helped us compile a list of relevant domain keywords. We then assessed various combinations of these keywords to determine their effectiveness in retrieving a collection of studies. Consequently, three query strings were employed to identify primary studies, as illustrated in Table 4.

We chose several digital libraries for our research, including ACM Digital Library,[1] Scopus,[2] Springer,[3] ProQuest,[4] IEEE-Xplore,[5] and ScienceDirect,[6] due to their comprehensive coverage of leading journals and conference proceedings. Additionally, we conducted a search on Google Scholar[7] to enhance our research scope. In total, we used seven different databases in this study.

We conducted our final search using a timeframe restricted to the period from July 2010 to April 2024. The timeframe was set due to a significant increase in research papers about MT and formality transfer after 2010, compared to the period before. Based on our search method, we found a total of 145 primary studies. The detailed search process and the count of papers identified at each stage are illustrated in Fig. 4.

B. STUDY SELECTION

The preliminary search conducted through database queries yielded 2,630 possible papers (see Fig. 4). Table 5 presents the outcomes of this stage for each database along with the respective query strings. Then, we filtered out articles that were not relevant by reviewing the titles. To address result overlap from multiple digital libraries, we implemented a structured deduplication process to ensure each article was uniquely represented. This process involved two main steps: first, we compared titles for filtering out duplicates. Then, we manually reviewed a sample of the deduplicated dataset to verify accuracy and confirm that no important articles were excluded. As a result of this phase, 141 papers were obtained.

After that, we utilized the inclusion and exclusion criteria to select the most pertinent papers for our study. Papers that fulfilled all inclusion criteria were included, while those that met any exclusion criteria were not considered. In this step, we selected 77 papers by reviewing their abstracts and entire

[1] https://dl.acm.org
[2] https://www.scopus.com
[3] https://link.springer.com
[4] https://www.proquest.com
[5] https://ieeexplore.ieee.org
[6] https://www.sciencedirect.com
[7] https://scholar.google.com

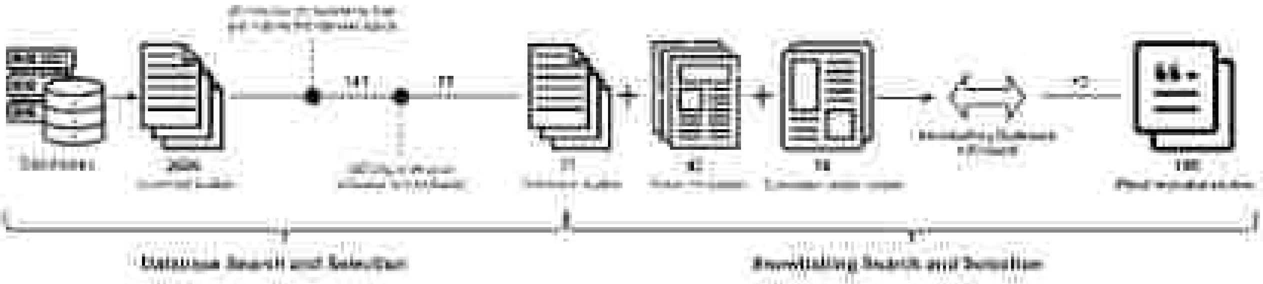

FIGURE 4. Search strategy and selection process.

TABLE 5. Results of preliminary database query search.

Database	Query 1	Query 2	Query 3
ACM	585	4	0
Scopus	79	16	0
Springer	69	12	0
ProQuest	101	25	0
ScienceDirect	5	1	0
SemanticScholar	93	3	0
Google Scholar	1430	162	1
Total	2362	219	1

texts. The following inclusion and exclusion criteria were used in our study.

- Inclusion criteria
    1) The paper should be written in English, as it is the common language used within the natural language processing community.
    2) The paper introduces a corpus that can be utilized for examining the formality transfer in Arabic text.
    3) The paper introduces an automatic evaluation metric that could be applied to Arabic text formality transfer.
    4) The paper describes a model or technique employed for MT or text formality transfer.
- Exclusion criteria
    1) Papers that fail to meet any of the outlined inclusion criteria.
    2) Extended abstracts, posters, books, patents, and tutorials.
    3) Papers do not concentrate on textual mobility.
    4) Papers do not employ formality attribute in text style transfer.
    5) Papers provide dataset that is not organized at the level of individual sentences.

For implementing the second phase of our hybrid search approach, we utilized the snowballing search to ensure that no pertinent studies were overlooked. As a result of applying snowballing search, we selected 77 papers that present techniques for MT or text formality transfer. Additionally, 47 papers have been selected for presenting datasets that can be utilized with Arabic text formality transfer. Moreover, we have chosen 18 papers that introduce automated evaluation metrics for assessing the performance of Arabic text formality transfer. It is worth mention here that two more papers are chosen because they provide both methodology and resources for text formality transfer. While, the final paper selected presents methodology and a new evaluation metric for text formality transfer.

Therefore, the chosen 145 papers introduce methodologies, resources, and assessment criteria for implementing text style transfer and MT. Fig. 5 visually represents the scope of our work by displaying the top 10 most frequent words found in the titles of the selected papers. Whereas, Fig. 6 provides additional insight into our research by visually presenting the 30 most common 2-gram combinations (adjacent word pairs) extracted from the titles of the selected papers.

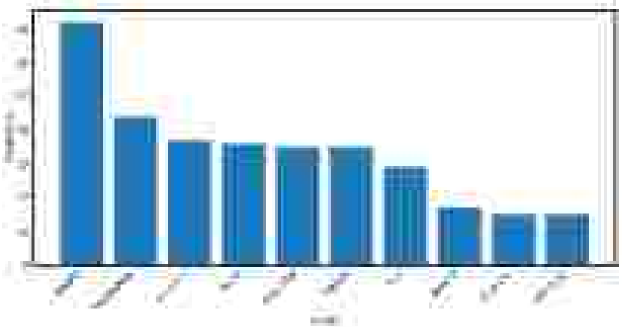

FIGURE 5. The 10 most commonly occurring words in the titles of the 145 papers.

The final studies chosen were based on the quality assessment criteria we developed to evaluate the relevance and strength of the primary research. The criteria for quality assessment can be found in Table 6. We focused exclusively on studies that had a quality score exceeding 50% of the highest possible score, and these were the ones selected for data extraction.

C. DATA EXTRACTION AND ANALYSIS

We gathered pertinent data from each chosen paper to address the RQs. Additionally, we compiled metadata for each paper to support further statistical analysis. This metadata comprised the title, year of publication, authors,

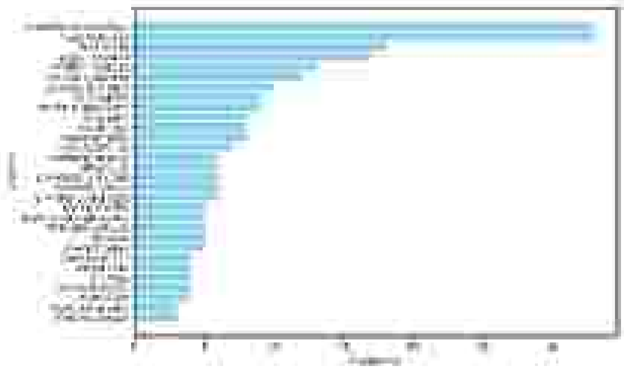

FIGURE 6. The 30 most 3-gram combinations found in the titles of the 145 papers.

TABLE 6. Quality assessment questions.

QA	Quality question
1	Is the methodology of the paper clearly defined?
2	Has the dataset size and data source been clearly identified?
3	Are the machine learning techniques adequately described?
4	Does the research make a meaningful contribution?
5	Can the presented methodology be applied and evaluated?

and publication type. The collected data were systematically arranged in Excel spreadsheets.

The main objective of data synthesis is to gather and integrate information and statistics from chosen studies to address the RQs and formulate a response. By grouping studies with similar and comparable results, it was possible to provide definitive answers to the RQs through the presentation of research evidence.

We analyzed a range of data, including numerical performance metrics, categorization of approaches, machine learning methods, languages, domains, and datasets. To integrate information from the main studies and tackle the RQs, we employed different methods, including visualization tools. Additionally, tables were utilized to summarize and convey the results.

IV. RESULTS

In this section, we provide an in-depth summary of the results from our literature review, addressing each RQ in detail.

A. RESOURCES (RQ1)

Several resources [61], [62] are available for gathering data used in Arabic text formality transfer. Some of them are designed for particular Arabic dialects [63], while others are suitable for use across multiple dialects and MSA. The following subsections detail these resources.

1) PARALLEL DATASETS

In this subsection, we offer a survey of prevalent parallel datasets [64] for Arabic dialectal MT as documented in existing literature. A summarized view of these datasets, specifically tailored for translation between ADs and MSA, is presented in Table 7. The following subsections provide detailed explanation into the particularities of every dataset presented.

TABLE 7. Summary of parallel public datasets for ADs and MSA.

Dataset	Dialects	Samples #	Year
APT	Multiple dialects	MSA-Eng sentences: 8.5M; Levantine-Eng sentences: 138K; Egyptian-Eng sentences: 38K	2012
AAD	Algerian dialects	4K sentences	2014
MDPC	Five dialects	1K sentences; 2K each AD and MSA/Eng	2014
PADIC	Multiple dialects	6.4K sentences	2015
MADAR	25 cities dialects	2K sentences each 25 ADs; 12k sentences each 5 ADs and MSA	2018
[illegible]	Multiple dialects	[illegible]	2018
[illegible]	Multiple dialects	[illegible]	2019
[illegible]	[illegible]	[illegible]	2018
[illegible]	Moroccan, Tunisian	[illegible]	2019
[illegible]	[illegible]	[illegible]	2020
[illegible]	Tunisian	[illegible]	2020
[illegible]	Egyptian, Saudi	[illegible]	2017
Lisan	Multiple dialects	1.2 million tokens	2021
Nadia	Syrian dialect	5K sentences	2020
NADI	Multiple dialects	[illegible]	2021

a: ARABIC-DIALECT/ENGLISH PARALLEL TEXT (APT) CORPUS

Zbib et al. [65] introduced a dataset containing Arabic-English translations, including MSA and two ADs. The dataset is made up of over 8 million sentences in MSA and English, along with 138,000 sentences in Levantine Arabic and English, and 38,000 sentences in Egyptian Arabic and English. The corpus was sourced from Arabic weblogs, and the conversion process was conducted using Amazon Mechanical Turk. The dataset was collected to be a valuable asset for training and evaluating MT models, particularly for handling ADs and MSA.

b: ALGERIAN ARABIC DIALECTS (AAD)

Smaili et al. [66] aim to create parallel corpora for Algerian dialects for the first time. Their ultimate goal is to facilitate MT between MSA and Algerian dialects. Additionally, they introduce language tools designed to handle these dialects. Initially, they adapted BAMA, a renowned MSA analyzer, to develop a morphological analysis model for the dialects. Following this, they propose a diacritization system based on an MT process to reinsert vowels into dialect corpora.

c: MULTI-DIALECTAL PARALLEL CORPUS (MDPC)

Bouamor et al. [67] developed this corpus by extracting two thousand Egyptian-English sentences out of APT dataset, originally created by Zbib et al. [65]. Following the selection process, indigenous speakers originating from Palestine, Syria, Jordan, and Tunisia were tasked with translating the sentences into their respective native dialects. They also translated an additional 8000 sentences from Egyptian-English data specifically into Syrian dialect. This dataset provides a valuable resource for comparing identical sentences translated into different dialects.

d) PARALLEL ARABIC DIALECT CORPUS (PADIC)

The PADIC is a notable resource in the realm of Arabic dialectal studies and NLP. Meftouh et al. [68] present this corpus, which is a multi-dialect dataset comprising of MSA, Algerian, Tunisian, Palestinian, and Syrian dialects. This corpus encompasses 6.4K parallel sentences for MSA and its dialects. It was generated from manually transcribed recordings of daily discussions, as well as scripts from movies and TV dialogues, which were converted into six varieties. It consists of over 37,000 tokens, with approximately 10,000-word types in both MSA and the five dialects.The corpus is meticulously assembled to incorporate a broad spectrum of linguistic phenomena, expressions, and idiomatic usage unique to each dialect. Its accessibility has expedited [69] the creation of algorithms, models, and utilities designed to accommodate the distinctive features of Arabic dialects.

e) MULTI-ARABIC DIALECT APPLICATIONS AND RESOURCES CORPUS (MADAR)

The MADAR dataset constitutes an extensive compilation of Arabic text data representing a range of dialects prevalent across the Arab region. It encompasses texts sourced from diverse platforms like news articles, social media content, blogs, among others. The corpus is meticulously annotated and categorized, facilitating research endeavors and advancements in NLP, machine learning, and computational linguistics pertaining to Arabic dialects.

Bouamor et al. [70] curated 2k English sentences from the BTEC corpus, initially a Japanese/English repository for parallel phrases, and engaged native speakers from 25 Arab cities to translate these phrases into their respective dialects. They further translated 10k sentences for five specific cities: Doha, Beirut, Cairo, Tunisia, and Rabat. The initial resource comprises a substantial parallel dataset encompassing 25 Arabic city dialects, along with the prior parallel sets of English, French, and MSA dialects. While the second corpus consists of a lexicon with entries in the dialects of 25 cities, totaling 1,045 entries each, accompanied by equivalents in MSA, French, and English.

f) Dial2MSA

Mubarak [71] introduced a parallel dataset for translating Arabic dialect to MSA. This corpus includes 5,000 tweets written in Egyptian/Maghrebi dialects and 6,000 tweets written in Levantine/Gulf dialects. Each tweet was translated to MSA by someone who speaks that specific dialect as their native language.

g) SAUDI DIALECT CORPUS (SUAR)

Al-Twairesh et al. [72] developed a dataset consisting of over hundred thousand words generated from various social media platforms including, X (formerly Twitter), YouTube, WhatsApp, blogs, Instagram and forum.

h) CITY-LEVEL DATASET OF ARABIC DIALECTS (CLDAD)

Mageed et al. [73] developed this massive Arabic dialect corpus collected from 10 Arabic countries, representing 19 cities containing varying dialects such as Egyptian, Gulf, KSA, Levantine, and Yemen. The data was gathered using the X API, employing multiple bounding boxes across several Arab countries.

i) THE BIBLE

The Bible corpus [74] contains Bible texts translated into multiple languages, serving as a valuable asset for MT endeavors. With translations available in numerous languages, it enables researchers and developers to train and assess MT models across various language combinations. The Bible has been translated into over 3300 languages. While there are numerous translations available in Arabic, translations into Arabic dialects remain extremely limited.

j) SADID

The SADID dataset [75] encompasses English, Egyptian, Levantine, and MSA. The dialectal materials were sourced from three different origins: 1) Wikipedia, chosen for its wide-ranging topics and straightforward language, 2) Aesop's Fables, selected for their narrative format, and 3) particular dialogues extracted from movie subtitles.

k) TUNISIAN ARABIC DIALECT (TAD)

Kchaou et al. [76] introduce a data augmentation method designed to generate a parallel corpus, aligning Tunisian Arabic dialect as expressed on social media with standard Arabic. This corpus aims to support the development of a MT system.

l) ARABIC SEMANTIC TEXTUAL SIMILARITY (STS)

The aim of Arabic STS dataset [77] is to ascertain the semantic similarity between two Arabic sentences. Each English phrase was translated into three distinct target languages: MSA, Egyptian Arabic, and Saudi dialect.

m) YEMENI, IRAQI, LIBYAN, AND SUDANESE ARABIC DIALECT CORPORA WITH MORPHOLOGICAL ANNOTATIONS (LISAN)

Jarrar et al. [78] present Lisan, a morphological annotated corpora collected from five Arabic cities dialects, including Yemeni, Iraqi, Libyan and Sudanese dialects consisting of about 1.2 million tokens. Specifically, Yemeni data which is about 1.05 million tokens was generated spontaneously from X platform, while the rest of the three cities (i.e. about 50K token each) was curated manually from Facebook and YouTube.

TABLE 8. Non-parallel corpora for AD sentences.

Reference	Year	Task	Dialect
[81]	2018	Multiple	Palestinian
[82]	2018	Multiple	Multi-dialectal
[83]	2017	Machine translation	Tunisian
[84]	2017	Multiple	Palestinian
[85]	2017	Sentiment analysis	Algerian
[86]	2018	Multiple	Multi-dialect
[87]	2018	Multiple	Multi-dialect
[88]	2019	Multiple	Multi-dialect
[87]	2020	Multiple	Maghrebi
[88]	2021	Multiple	Moroccan
[90]	2022	Multiple	Palestinian
[91]	2022	Sentiment analysis	Algerian
[92]	2022	Multiple	Multi-dialect
[93]	2024	Multiple	Multi-dialect
[94]	2024	Multiple	Yemeni

a: SYRIAN ARABIC DIALECTS WITH MORPHOLOGICAL ANNOTATIONS (NABRA)

Nayouf et al. [79] developed a Syrian Arabic dialect corpus with morphological annotations called Nabra. Nabra was curated by group of Syrian natives consisting of 6K sentences, totaling 60K words, from various outlets such as social media, movie and series scripts, song lyrics, and local proverbs.

b: THE KIND DATASET

Yamani et al. [80] presented a method of gathering data that is crafted to enable participants to translate sentences from MSA into their respective regional dialects. The MSA sentences are drawn from a subset of 11,670 sentences sourced from a widely recognized MSA dataset, specifically the MADAR dataset. Participants are tasked with translating these sentences into the dialect they identified as their own upon entering the competition. Another dataset for question answering is presented as well with 796 samples.

2) NON-PARALLEL CORPORA

Various non-parallel datasets are available for different NLP tasks, which can also serve as a source for implementing Arabic text formality transfer. Table 8 provides several examples and emphasizes that datasets from various tasks can also be leveraged in this line of research.

3) BENCHMARK DATASETS

A benchmark dataset is a standardized collection of data employed to assess and contrast the performance of models within a particular field. These datasets are crucial for gathering data applicable to the task of Arabic text formality transfer.

For example, Nagoudi et al. [95] introduce Dolphin, a groundbreaking benchmark designed to meet the requirements for evaluating NLG across various Arabic languages and dialects. This innovative benchmark covers an extensive array of 13 distinct NLG tasks, such as dialogue generation, question answering, machine translation, and summarization, among others. Dolphin includes a significant collection of 40 diverse and representative public datasets, organized into 50 test splits, meticulously selected to represent real-world contexts and the linguistic diversity of Arabic.

Similarly, Elmadany et al. [96] present ORCA, a publicly accessible benchmark designed for evaluating Arabic language understanding. ORCA has been meticulously developed to encompass various Arabic dialects and a broad spectrum of challenging comprehension tasks, utilizing 60 distinct datasets across seven natural language understanding (NLU) tasks.

4) SHARED TASKS

In a research community, a shared task usually involves a collaborative endeavor where various teams or individuals tackle the same problem or dataset to evaluate and compare their methods and outcomes. These collaborative challenges are particularly prevalent in the field of NLP. Such shared tasks can be employed to gather data for Arabic text formality transfer.

For example, Al-Khalifa et al. [97] proposed a shared task[8] for AD to MSA MT. Additionally, Abdul-Mageed et al. [98] proposed a shared tasks for AD-MSA MT. Moreover, we can also collect data from some related task such as AD identification [99], [100].

5) TOOLS

There are various tools and systems that can be used for collecting data and improving performance of Arabic text formality transfer, as explained in the following subsections. In this work, we specifically choose tools and systems that can be applied to both Arabic dialects and MSA.

a: MAGEAD

MAGEAD [101] is a tool designed for analyzing and generating morphological structures within the Arabic language family. It supports both MSA and various spoken dialects, initially focusing on the Levantine dialect. The system is adaptable to new dialects even without an existing lexicon, requiring only minimal manual knowledge engineering.

b: A HYBRID APPROACH FOR CONVERTING WRITTEN EGYPTIAN COLLOQUIAL DIALECT INTO DIACRITIZED ARABIC (HACWECDDA)

The HACWECDDA system [102] transforms a written sentence in Egyptian colloquial Arabic into a diacritized MSA sentence. Egyptian Colloquial Arabic was selected due to its widespread usage in blogs and articles online. Resources were gathered using a rule-based approach from a vast amount of data on the internet. The system's lexicon consists of 41,705 words, including 9,085 non-MSA words, 3,000 distinct colloquial words, with the remainder being spelling errors and non-Arabic names. The system is designed to differentiate between MSA and colloquial words.

[8]https://osact-lrec.github.io/

c: COLABA

Diab et al. [103] introduced COLABA, an extensive initiative aimed at developing resources and tools for processing Dialectal Arabic. Their efforts focus on enhancing the existing resources and tools, expanding them to cover additional dialects and varieties of dialectal data.

d: ARABIC DIALECT HANDLING IN HYBRID MACHINE TRANSLATION (ADHHMT)

The ADHHMT system [104] expands on a hybrid MT system designed for managing Arabic dialects. It incorporates a Statistical decoder featuring four rule categories: lexical, syntactic, argument structure, and functional structure rules, alongside semantic disambiguation data, a statistical bilingual lexicon, bilingual phrase table, and target language models.

e: DIALECTAL TO STANDARD ARABIC PARAPHRASING TO IMPROVE ARABIC-ENGLISH STATISTICAL MACHINE TRANSLATION (DSAPAESMT)

The DSAPAESMT system [105] generates standardized paraphrases in MSA. It incorporates two dialect varieties: Levantine and Egyptian. A simple rule-based method is applied with this system. An existing MSA analyzer is enhanced by including dialect-specific out-of-vocabulary (OOV) words and low-frequency words.

f: THE ARABIC ONLINE COMMENTARY DATASET (AOCD)

This AOCD system [106] utilizes a vast monolingual dataset of 52 million words, emphasizing diverse dialectal content. It is specifically trained to detect and quantify dialectal elements within sentences. The data is sourced from three newspapers renowned for their extensive use of Levantine, Gulf, and Egyptian dialects. Moreover, the system can differentiate between dialectal content and MSA, as well as between various dialectal forms.

g: SENTENCE LEVEL DIALECT IDENTIFICATION FOR MACHINE TRANSLATION SYSTEM SELECTION (SLDIMTSS)

The SLDIMTSS system [107] can identify whether a sentence is purely MSA or contains dialectal elements, enabling the use of the most appropriate MT system for better accuracy.

h: MULTI-DIALECT MACHINE TRANSLATION (MUDMAT)

The MuDMaT project [108] seeks to promote the advancement of MT systems tailored for under-resourced languages, including various dialects and variants. Specifically, it focuses on enhancing MT capabilities for three North African Arabic dialects—Tunisian, Algerian, and Moroccan—which face severe resource limitations. The project's methodologies and insights are transferable to other less-resourced language variants and dialects. The MuDMaT project focuses on developing hybrid MT systems that combine statistical and rule-based approaches. These systems aim to translate between various Arabic dialects, MSA, and French bidirectionally.

i: A NOOJ TUNISIAN DIALECT TRANSLATOR

Torjmen et al. [109] have developed a translator to convert the Tunisian dialect into MSA. The suggested translation approach relies on a bilingual dictionary derived from the studied corpus, supplemented by a comprehensive set of local grammars. These local grammars are then converted into finite state transducers using the latest technologies of the NooJ linguistic platform.

j: TOWARDS A PORTABLE SPOKEN LANGUAGE UNDERSTANDING SYSTEM APPLIED TO MSA AND LOW-RESOURCED ALGERIAN DIALECTS

To circumvent the need for extensive parallel corpora in MT, Lichouri et al. [110] address the challenge of interpreting user intents from natural language queries into a system's semantic format across various languages and dialects. Specifically, they focus on English, MSA, and four distinct Algerian vernacular dialects originating from Blida, Djelfa, Tenes, and Tizi-Ouzou regions.

k: A Seq2Seq NEURAL NETWORK BASED CONVERSATIONAL AGENT FOR GULF ARABIC DIALECT

Alshareef and Siddiqui [111] present an initial promising effort towards developing an open-domain conversational agent in the Gulf Arabic dialect. They employed a Seq2Seq neural network to construct a conversational agent using the Arabic Gulf dialect.

l: CAMeL TOOLS

The CAMeL tools [112] comprises a set of open-source Python tools designed for Arabic NLP. CAMeL Tools currently offers functionalities including text preprocessing, morphological analysis, dialect identification,[9] named entity recognition, and sentiment analysis.

m: BUILDING THE EMIRATI ARABIC FrameNet (EAFN)

The EAFN project [113] seeks to establish a FrameNet specifically for Emirati Arabic, leveraging the Emirati Arabic Corpus. The objective is to develop a resource akin to the early stages of the Berkeley FrameNet. The project is structured into manual and automatic tracks, reflecting the main methods employed to gather frames in each track.

n: ONLINE TOOLS

There are several online tools available that can be used for transforming the formality of Arabic text. For example, the AEADAIT system is an AI-driven tool[10] crafted for precise dialectal translation, cultural understanding, and educational assistance.

[9]https://camel-tools.readthedocs.io/en/latest/api/dialectid.html
[10]https://www.yeschat.ai/gpts-[illegible]-Arabic-Explorer

B. EVALUATION METRICS (RQ2)

It is difficult to design reliable scoring systems [114] for MT because there is not always a single "correct" way to translate a sentence. One method for assessing MT involves human evaluation [115]. Nonetheless, the need for automatic evaluation metrics remains significant in this area of research. Existing metrics [116] mostly judge translations by how closely the words match a reference translation, with the main difference being how they handle word order changes and synonyms. This subsection discusses common metrics employed in literature to automatically evaluate the performance of MT systems.

In this study, we present the automation evaluation since it standard with this research work. It is important to note that while some research includes human evaluation, we have chosen not to incorporate it in our study. This is due to the lack of standardization and the susceptibility to human errors.

1) METEOR

METEOR, a popular tool used in MT and other NLP tasks, compares draft translations to corrected ones. It considers individual words (unigrams) and uses three (alpha, beta, and gamma) to adjust its scoring based on the specific language and type of translation being done. Unlike simpler methods, METEOR can handle situations where words are rearranged or synonyms are used, making it valuable for assessing MT quality across various languages and situations.

METEOR developed by Lavie and Denkowski [117], is a performance measure employed to assess the efficiency of MT systems by comparing individual words in a generated translation to their corresponding words in a reference translation. When a word in the candidate translation has multiple matches in the reference translation, METEOR favors alignments that require the least shuffling of words. The extent of rearrangement is determined by tallying the number of crossed alignments. The final alignment is used to compute the unigram precision and recall of the machine translation given by equations 1 and 2.

$$Presion = Matches/(Length\ trans) \tag{1}$$

$$Recall = Matches/(Length\ ref) \tag{2}$$

The precision and recall are exploited to obtain the weighted unigram harmonic mean illustrated by equation 3. Equation 4 presents the final equation for METEOR, which is a multiplication between the unigram harmonic mean F_α and the penalty obtain from equation 5. METEOR includes three settings (α, β, and γ) that can be adjusted to better match human evaluation of translations for a particular language and type of translation task.

$$F_\alpha = Presion/(\alpha Presion + (1-\alpha)Recall) \tag{3}$$

$$METEOR_{\alpha,\beta,\gamma} = F_\alpha \cdot P_{\beta,\gamma} \tag{4}$$

$$P_{\beta/\alpha} = 1 - \gamma\left(\frac{Chuncks}{Matches}\right)^{\beta} \tag{5}$$

2) BILINGUAL EVALUATION UNDERSTUDY (BLEU)

The BLEU metric was introduced by Papineni et al. [118]. It is a common metric for checking the quality of MT, especially in the world of language processing. It works by looking for how many words match between the machine translated output and the human-made generated translation. BLEU considers groups of words (n-grams) and rewards translations with more overlap. However, it also discourages translations that are too short, ensuring they capture the full meaning of the original text. It penalizes translations that are not longer than the reference by adjusting the final score based on the length difference.

Equation 6 shows how BLEU calculates a score for machine translation using matching word groups (n-grams) up to a certain span (N). This equation considers a collection of machine generated translations (T) and their corresponding human-made translations (R). Whereas, BP value is calculated by using equation 7.

$$BLEU\ (N) = \left(\prod_{n=1}^{N} \frac{n-grams\ (T \cap R)}{n-grams\ (T)}\right)^{\frac{1}{N}} BP \tag{6}$$

$$BP = min\left(1.0, e^{1-len(R)/len(T)}\right) \tag{7}$$

BLEU is easy to use, but it has some weaknesses. For example, BLEU does not care about the order of words. As long as the translation includes the right groups of words (n-grams), it can get a good score even if the sentence structure is messy. BLEU also does not check if any important information is missing from the translation. It only looks for matching n-grams, not whether the translation captures everything from the original text. Finally, BLEU can not tell the difference between words or phrases with similar meanings. Despite this limitation, BLEU remained one of the most employed metric for assessing the precision of different MT systems.

It is worth mention here that some studies assess the BLEU score by comparing transferred sequences to the original sequences [119], whereas others evaluate the BLEU score by comparing the transferred sequences to human [120].

3) TRANSLATION EDIT RATE (TER)

TER was proposed by Snover et al. [121] to establish the quickest way to convert candidate translation by correlating it to a perfect translation done by a human. It can insert, delete, or swap words, or even switch short phrases around. This "swap" ability is what makes TER different from a trivial performance measure called word error rate (WER), which can only insert, delete, or swap individual words. By allowing swaps, TER avoids unfairly penalizing translations where the word order is slightly different. Specifically, TER score is obtained based on equation 8. From a MT standpoint, a lower TER score means the MT is closer to the human-made version.

$$TER = \frac{(Num\ edits)}{(Average\ ref\ len)} \tag{8}$$

Snover et al. [122] proposed TER-Plus (TERp), which is basically an improvement on TER. It lets researchers adjust the importance of different editing actions (like insertions or deletions) to better match how humans judge translation quality. Additionally, TER-Plus can consider extra ways to fix translations, including word stem matches, WordNet synonym and multiword matches.

4) CROSS-LINGUAL OPUS METRIC MEASURES TRANSLATION (COMET)

Rei et al. [123] developed COMET, which is among the most popular performance metrics for evaluating MT over the past few years. It focuses on how well the translation matches the original text word-for-word, across different languages. It does this by comparing the words in the original text and the translation one by one. This approach helps assess how accurate and faithful the translation is, especially when working with different languages.

5) CHARACTER n-gram F-SCORE (ChrF AND ChrF++)

ChrF and ChrF++, proposed by Popovic [124], are performance measures developed to assess the performance of MT algorithms. They fall into a class of metrics that concentrate on similarities at the character level between the candidate translation and perfect translations performed by human. ChrF (see equation 9) checks how similar the character n-grams between machine-generated translations and reference translations by employing both precision and recall for calculating an F-score that harmonizes those aspects.

This metric is valuable for assessing translations in scenarios with ambiguous word boundaries or languages with intricate morphology, were conventional word-based metrics encounter limitations. ChrF considers how many of those character sequences match exactly (precision) and how many do not get missed (recall) and combines these aspects into a single score. ChrF++ extends the concept of ChrF by considering word order up to 2.

$$ChrF\beta = \left(1+\beta^2\right)\frac{CHRP \times CHRR}{\beta^2 \times CHRP + CHRR} \tag{9}$$

where $CHRP$ and $CHRR$ represent the mean precision and recall values of character n-grams, respectively, encompassing all n-grams. Similar β denotes the weights.

6) BERTScore

Zhang et al. [125] introduced BERTScore as a method that goes beyond merely analyzing single words. It uses a powerful language model called BERT and its contextual embedding to grasp the overall interpretation of the sentence. By comparing the generated MT to a perfect translation, BERTScore can give a similarity score that reflects how well the translation captures the original meaning. Mathematically, BERTScore can be calculated based on equation 10.

$$BERTScore = \frac{1}{|C|}\sum_{c \in C}\max_{r \in R}\cos\left(e_c, e_r\right) \tag{10}$$

where, $|C|$ denotes the amount of tokens from the machine generated translations, while C is the sets of tokens of the machine generated translation. Similarly R denotes the token embeddings of the human perfect translations, while e_c and e_r denotes the embedding of token c and r, respectively.

7) RECALL-ORIENTED UNDERSTUDY FOR GISTING EVALUATION (ROUGE)

ROUGE score is centers on assessing the extent of content similarity between the machine generated translation and reference translations. This evaluation measure can be calculated by using equation 11 [126].

$$ROUGE = \frac{\sum_{n\text{-grams}} \text{Recall}_{n\text{-grams}}}{\text{Total n-grams in ref. translations}} \tag{11}$$

8) PERPLEXITY (PPL)

Perplexity (PPL) score [127] gauges the level of astonishment the model experiences when confronted with unfamiliar data. It is used for a fluency evaluation. Lower perplexity values indicate more effective training. This metric can be calculated by using equation 12.

$$PPL = \exp\left(-\frac{1}{N}\sum_{i=1}^{N}\log p(x_i)\right) \tag{12}$$

where N represents the count of tokens within the same translated sentence, x_i denotes every token within the same translated sentence, and $p(x_i)$ represents the likelihood attributed to each token x_i by the probability distribution.

9) OTHER METRICS

Other metrics are also utilized in some related studies, though they are not as commonly employed. For example, the Generalized Language Evaluation Understanding (GLEU) metric [128], inspired by BLEU, serves as a tool to assess the preservation of content. There is also a measure for style accuracy (SAcc) by evaluating the prediction accuracy of a pre-trained style classifier on the generated sentences. Additionally, the context coherence can be evaluated using a method described by Cheng et al. [129]. This approach involves using a pre-trained coherence classifier to predict how well a generated sentence aligns with its surrounding context, based on the prediction accuracy.

Additional metric entitled as Paraphrase In N-gram Changes (PINC) [130] is used also with this research direction. PINC indicates the degree of difference from the original informal sentences, but this does not necessarily correspond to the effectiveness of the style transfer. Alternative metrics encompass Cosine Similarity (CS) [131], which assesses similarity between the embeddings of original and style-transferred sentences to evaluate content preservation. Similarly, the degree of Meaning Similarity (MS) [132] between the original and modified texts is evaluated using the SentenceTransformers embedding distance. In a similar context, Hamming distance (Eh) calculates the variance between the source and generated texts. Moreover, Word Overlap (WO) [133] measures the unigram overlap rate between the original and style-altered sentences, while BARTScore [134] evaluates the generated text from multiple perspectives, including informativeness and fluency. In the same context, Sellam et al. [135] introduced BLEURT as a score for assessing the similarity between translation candidates and reference texts. Unlike methods that rely on N-gram overlap, BLEURT is capable of recognizing semantic differences.

Certain studies substitute PPL with the precision of a RoBERTa-large model that has been trained on the CoLA corpus [136] for evaluating grammatical test (GT). Similarly, Mir et al. [137] suggested an alternative approach to assessing fluency, proposing the use of classifiers trained to differentiate between machine-generated and human-written sentences, rather than relying on perplexity.

Moreover, certain studies assess semantic similarity (SIM) by employing the unbiased embedding-based SIM model introduced by Wieting et al. [138]. This metric shows strong performance on semantic textual similarity (STS) benchmarks in SemEval evaluations [139].

Some studies employ the Geometric Mean (GM) to compute the geometric mean of accuracy, BLEU, and 1/PPL. Other research presents findings by using the harmonic mean (HM) of BLEU and style accuracy as a comprehensive score. However, we do not include these metrics in our analysis since they only reflect general performance.

C. COMPARISON BETWEEN EVALUATION METRICS

The metrics can be categorized according to the technique they use into three groups: "N-gram-Based Metrics", "Semantic-Based Metrics", and "Edit-Based along with Miscellaneous Metrics". Fig. 7 shows the category to which each metric belongs.

FIGURE 7. Categorization of evaluation metrics by technique.

It is clear that BLEU, ChrF, and TER mainly focus on comparing individual words or characters between the candidate translation and perfect translation done by human. COMET, on the other hand, takes a different approach. It uses a special technique to grasp the meaning of the words (semantics) and check if the translation captures the same meaning as the original text. Therefore, we can also categorize evaluation metrics based on their intended purpose into three groups: "Fluency and Adequacy", "Semantic Similarity and Alignment", and "Model and Data Properties" as illustrated in Fig. 8.

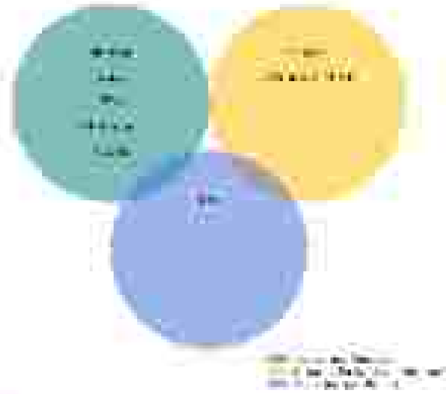

FIGURE 8. Classification of evaluation metrics based on purpose of use.

D. TAXONOMY OF APPROACHES FOR FORMALITY TRANSFER

In this study, the various methodologies employed in the chosen papers are classified to enhance the value of the research for future scholars. Fig. 9 presents a taxonomy of reviewed approaches alongside selected papers. This classification is developed according to the methods used, the data configurations, and the significant results achieved through different approaches.

The figure illustrates the classification of approaches into three main categories: conventional, seq2seq, and LLMs. We distinguish LLMs separately to highlight their significant role due to recent advancements in NLP. Conventional methods are further categorized into rule-based, statistical, and hybrid techniques. The seq2seq approaches are classified into supervised, unsupervised, semi-supervised, and reinforcement learning techniques. Within unsupervised learning (Non-parallel Data), there are four techniques employed: disentanglement, prototype editing, back-translation/pseudo data creation, and paraphrasing [140].

E. ARABIC TEXT FORMALITY TRANSFER (RQ1)

In this section we explored various studies in the context of Arabic formality transfer ranging from classical approaches, neural machine translation techniques (i.e., sequence to sequence methods) and the recent much anticipated LLMs.

We explored previous studies concentrated on the translation from ADs to MSA and vice versa, as well as between

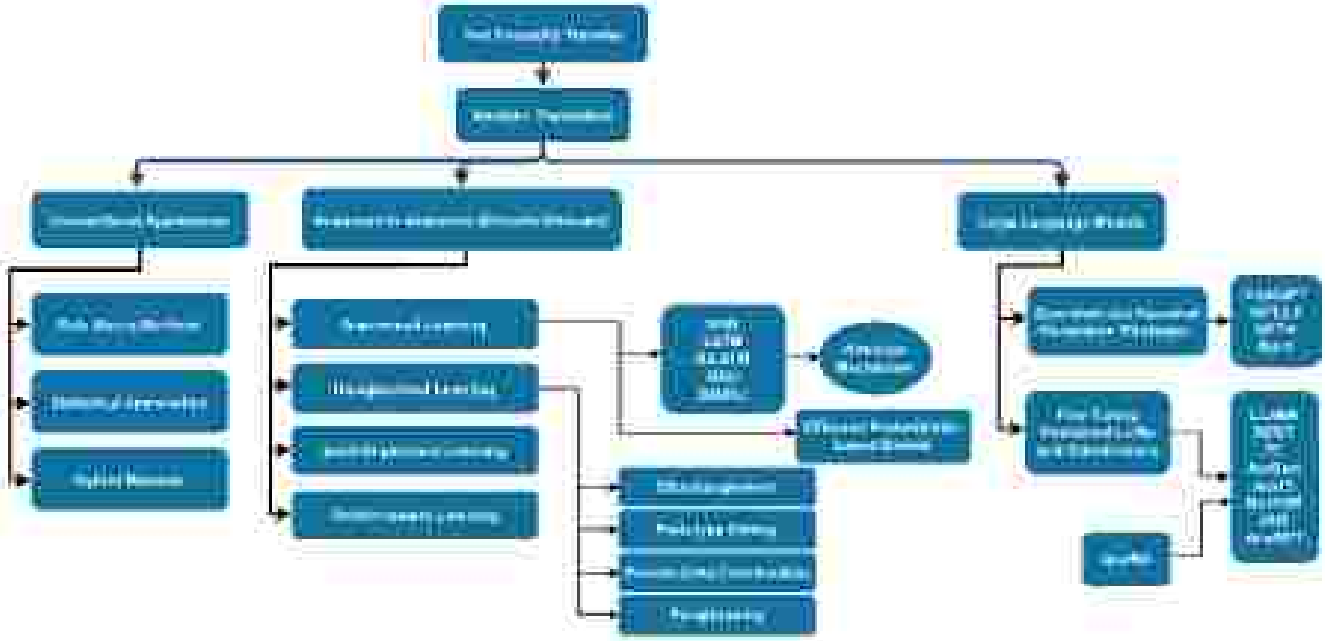

FIGURE 9. **A taxonomic analysis of approaches in light of selected papers.**

MSA and English in both directions [21]. Additionally, recent significant works have been discussed as a NMT, ranging from the utilization of basic sequence-to-sequence models to the integration of self-attention mechanisms, thereby establishing transformers and extensive language models as the prevailing approach in NMT. Furthermore, recent advancements in text style transfer are discussed, emphasizing their utility in enhancing the Arabic text formality transfer.

1) CONVENTIONAL APPROACHES

There has been increased interest in research on translating low-resource languages like Arabic (i.e., translation between AD and MSA) over the last few years. Thereby, a lot of research has been observed in the context of ADs translation [141].

A lot of the reviewed papers focus on the conventional approaches [142], [143], [144] such as rule based [68], [145], [146], [147], [148], [149], [150], [151], [152], statistical approaches [153], [154], [155], [156] and hybrid methods [104], [157], [158]. For rule-based MT, a set of rules based on the grammar and vocabulary of both the source and target languages is used to translate text. Specifically, it offers variety of approaches including direct translation, transfer system and interlingual system. While, statistical translation computes the probability $P(S/T)$ for any given source candidate S and target candidate T, selecting the translation that maximizes the probability.

Al-Gaphari and Al-Yadoumi [150] employed rule-based approach for translation between Sana'ani dialect and MSA with achieving 77.32 % performance. Similarly, Tachicart and Bouzoubaa [157] employed rule-based technique that depends on language models for transforming the Moroccan dialect to MSA.

The rule-based MT techniques have significant drawbacks. For example, constructing rule-based demands considerable amount of time. Whereas, statistical methods require high computational devices, and they cannot address syntactic challenge present in Arabic dialect such as word ordering. In general, the conventional methods often produce conflicting sentences that are incompatible or linguistically inconsistent. Fig. 10 provides an overview of the MT systems employed for translating Arabic dialects to MS using conventional approaches. It is evident that a rule-based approach is more commonly employed than a statistical approach in the reviewed studies.

2) SEQUENCE TO SEQUENCE APPROACHES

Despite the fact that research on translating AD to MSA and vice versa is dominated by the conventional approach like statistical, many researchers have started exploring the translation of AD to MSA based on NMT methods [159], [160], [161], [162], [163], [164], [165].

In particular, Baniata et al. [162] introduced the first NMT architecture designed to transform from ADs to MSA using RNN-based encoder-decoder architecture. Their method introduced a multi-task learning algorithm where the decoder is shared among every language combination, while each source language employs its own individual encoders. The results obtained have demonstrated that the developed model achieved high translation performance against the standalone models given a small training sample. Hence, multitask

FIGURE [illegible]. Summary of MT systems used with Arabic dialect to MSA translation based on conventional approaches.

learning can handle the challenge of insufficient training samples. Even though the developed system introduces an innovative concept of employing distinct encoders for every single language or dialect, along with a shared decoder. The proposed Arabic dialect-based translation system relies on the basic RNN encoder-decoder architecture.

In the same context, Guellil et al. [159] proposed a NMT model designed to process Arabic dialects, employing basic RNN encoder-decoder-based architecture. The system is specifically designed to translate between Algerian Arabic, a dialect that blends Arabic script with elements of Arabizi, and MSA. Similarly, Emna et al. [166] developed a NMT model to convert the spoken Tunisian Dialect (TD) into MSA. Their study has been expanded [167] by utilizing additional models.

Farhan et al. [160] introduce an unsupervised NMT-based approach for translating Arabic dialect. They developed two models for translating Arabic dialect to MSA. The first model is developed using standard attentional sequence to sequence model. While, the second model is based on Google NMT system. They firstly evaluated their approach in supervised setting using parallel corpus for Jordian to MSA translation before implementing the unsupervised settings. In the unsupervised setting, the highest BLEU score achieved is 32.14 by using the first architecture on Saudi-MSA corpus. However, their developed approach was evaluated using single Arabic dialect and a single evaluation metric. To effectively ascertain the efficacy of their developed approach it should be subjected to multiple ADs and more than one evaluation measure. Similarly, Al-Ibrahim and Duwairi [161] developed a machine translation approach for converting Jordanian Arabic vernaculars into MSA based on encoder-decoder deep learning algorithm that employs RNN model. However, they use a very small size corpus which ultimately hindered the performance of their model. In addition, they only focus on one dialect, leading to a shortage of NMT systems that encompass the complete range of dialects.

Baniata et al. [163] presented an algorithm for a unified multitask-based neural MT with Arabic dialect. The developed NMT algorithm exploit RNN encoder-decoder-based framework and integrates part-of-speech tags linguistic resource. The encoder is shared between Arabic dialect and MSA translation task achieving an improved BLEU score. Similarly, Baniata et al. [164] proposed a novel reverse positional encoding NMT scheme to convert Arabic dialect to MSA. The developed model scheme employed multihead attention (MHA) technique. The integration of the novel encoding scheme and application of sub-word units in the MHA block enhances the encoder's sublayer in the developed approach. These enhancements effectively capture all dependencies among the words in Arabic dialect input text. The outcomes obtained illustrate the ability of the developed approach to mitigate the open grammar issue of Arabic dialect sentences.

Baniata et al. [165] developed the first transformer based NMT algorithm for Arabic dialect exploiting subunits. The developed approach relies on the subunits and the shared vocabulary among the source (i.e. Arabic dialects) and the target candidate (i.e., MSA) to upgrade the conduct of the multi-head attention mechanism of the encoder. Experiments translating Maghrebi, Levantine, Nile, Gulf, and Iraqi dialects to MSA revealed an improvement in the developed method's performance when evaluated using the BLEU score metric. The experimental findings have confirmed that the transformer based NMT model can effectively capitalize on the subword units to enhance the translation performance.

Faheem et al. [168] developed a supervised, semi-supervised, and unsupervised machine translation architectures for translating Egyptian ADs to MSA. They employed both parallel and monolingual datasets to evaluate the developed schemes. Specifically, they developed three neural translations algorithms. Their first model utilizes a sequence-to-sequence architecture, capitalizing on the significant overlap in vocabulary between Egyptian dialect and MSA, for effective learning of the word embeddings. The second translation model employed transformer framework trained based on only monolingual dataset, while the third model utilizes both parallel and mono lingual datasets in the initial supervise learning stage and training phase, respectively. Results obtained demonstrate that semi-supervised approach produced the best BLEU score against the supervised and unsupervised approaches.

To investigate the spread of misinformation about Covid-19 and its impact on communities in Algeria, Slim et al. [159] explored the issue of translating Algerian vernacular to the context of Covid-19 related social media interactions. The primary goal of the study is to enhance the comprehension of Algerian vernacular messages concerning Covid-19. The developed approach starts by employing an LSTM algorithm to filter messages and detect comments related to Covid-19. Afterwards, the identified Covid-texts are translated using embedded GRU model from Algerian vernaculars to MSA, achieving a BLEU score of 22.10.

In addition, Slim et al. [170] developed sequence-to-sequence algorithm using attention network to transform Algerian Arabic dialect to MSA. The developed algorithm is based on transductive transfer learning approach. The main objective for integrating transductive learning scheme is to address the learning problem encountered for under-resourced languages such as Arabic dialects (i.e., Algerian dialect). The training process began with a parallel dataset encompassing multiple Arabic dialects. Subsequently, the models were fine-tuned on a under-resourced data focused exclusively on the Algerian dialect. The impact of the transductive learning technique was highly noticed. Specifically, it greatly enhances the BLEU score accuracy of the sequence-to-sequence algorithm, increasing it from 0.3 to above 34. With the attentional based method, it increases the BLEU score from less than 17 to more than 35, thus demonstrating the capability of their proposed method.

Moreover, Al-Khanuri and AAlAbdulsalam [171] addressed the lack of specific research on MT between Omani Arabic dialect and English by developing an Omani dialect parallel dataset generated from social media source X. We chose to include this study in our research because the authors use transfer learning techniques for the Omani Arabic dialect based on a pre-existing MSA-English model. Consequently, this research can be viewed as a form of indirect translation between AD and MSA.

Moukafih et al. [172] have also tried various multitasking learning methods to leverage the connections among Arabic dialects. Table 9 shows a summary of sequence to sequence based NMT approaches for translation between AD and MSA.

3) LARGE LANGUAGE MODELS

LLMs are advanced pre-trained neural networks developed to comprehend and produce human-like text. These models constitute a substantial progress in NLP as they possess the capability to understand and generate text across different tasks like MT, question answering, text generation, etc. LLMs usually trained on extensive text datasets containing billions of words or even more, aim to grasp the underlying patterns and structures of human language. They utilize deep learning architectures like transformers, enabling them to capture extensive contextual information and dependencies within text sequences.

LLMs possess the remarkable capability to undertake diverse NLP tasks, eliminating the need for task-specific fine-tuning. Once trained on extensive text corpora, LLMs can be readily deployed for various downstream tasks with minimal additional training. Nonetheless, obstacles remain, including the shortage of annotated data for dialects, managing diverse dialectal variations, and preserving the cultural nuances inherent in each dialect. Famous examples of LLMs include OpenAI's GPT series, encompassing models like GPT-2, chat-GPT, GPT-3, and their succeeding versions, alongside models like bidirectional encoder representations from transformers (BERT).

Recently, researchers have started investigating the performance of LMMs in various NLP problems such as NLG subtasks. For instance, in the context of MT (such as translation of ADs to MSA), the process typically involves fine-tuning LLMs on dialectal data and MSA parallel corpora to improve their performance in understanding and generating accurate translations. The LLMs models can be exploited directly using the dialectal/MSA data via N-shot (also called in-context learning) settings. Table 10 illustrates a summary of MT systems used to convert Arabic dialect to MSA based on LLMs with ZS generation strategy and fine-tuning.

Recent works show that pre-trained LLaMA models, primarily trained on English-dominated corpora, inherently lack proficiency in handling non-English languages. To investigate this issue, Zhao et al. [180] seek to effectively adapt language generation and instruction-following skills for a language with limited resources. In particular, they conducted an extensive empirical investigation based on LLaMA to examine the need for expanding vocabulary and the necessary scale of training (i.e., additional pre-training, instruction tuning) for successful transfer. Their findings reveal that extending the vocabulary is unnecessary, and comparable transfer accuracy to advanced models can be attained with small additional pre-training samples in the majority of these tasks.

Alyafeai et al. [53] comprehensively evaluate the capability of two LLMs series (i.e., GPT3.5 and GPT-4) across seven different Arabic NLP tasks. The evaluation is carried out to gauge the abilities of these emerging foundational models and to compare their performance with advanced approaches in the literature. The extensive experiments indicate that GPT-4 surpassed GPT-3.5 in the majority of the 7 problems analyzed. The findings underscore a notable performance gap between the ChatGPT models and their Arabic-specific counterparts across most tasks. Both ChatGPT models exhibit exceptional performance when compared to current leading models. Nevertheless, certain limitations were noted in particular scenarios, such as summarization, MT, mostly for low-resource languages, and reasoning capabilities.

Ghaddar et al. [181] take a fresh look at how to pre-train and evaluate Arabic pre-trained language models (PLMs) by developing five new models based on BERT and T5 pre-training approach, achieving state-of-the-art performance against existing Arabic models on text generation task. Regarding pre-training, the authors investigate enhancing Arabic language models based on three features including,

TABLE 5. Summary of sequence to sequence based NMT approaches for translation between AD and MSA.

Reference	Year	Model	Dialects Considered	Evaluation Metric	Parallel Dataset	Result
[177]	2023	Neural Machine Translation	One Dialect (Algerian dialect)	BLEU	PADIC	[illegible]
[162]	2016	RNN-based encoder-decoder	Two dialects (Levantine(LA) and Maghrebi(MGR))	BLEU	PADIC and MPCA	LA-MSA: [illegible] MA-MSA: [illegible] MSA-LEV: [illegible]
[163]	2018	Multitask RNN encoder-decoder model with POS tagging	Levantine(LA) and Maghrebi(MA)	BLEU	PADIC and MPCA	[illegible]
[160]	2020	[illegible]	Multiple dialects (Jordanian, Saudi and Egyptian)	BLEU	Self created dataset	Supervised - [illegible] Unsupervised - [illegible]
[161]	2021	RNN encoder-decoder model	Single dialect (Sudanese dialect)	[illegible] and BLEU	Manually created dataset	Word level - [illegible] Sentence level - [illegible]
[152]	2022	Multitasking learning with GRU	Multiple dialects (Algerian (ALG), Moroccan (MAR), Tunisian (TUN), Syrian (SYR), and Palestinian (PAL))	BLEU	PADIC	[illegible]
[153]	2022	Transformer-based NMT	Multiple dialects (Levantine(LA), Maghrebi(MGR), Nile Basin Arabic, Gulf Arabic dialect and Iraqi Arabic dialect)	BLEU	PADIC, MPCA and MADAR	Transformer-based NMT improved model - [illegible]
[157]	2022	Embedding-GRU model	Single dialect (Algerian dialect)	BLEU	PADIC and [illegible]	[illegible]
[166]	2022	LSTM-GRU, [illegible] NMT and Universal transformer	Single dialect (Tunisian dialect)	BLEU	Self collected dataset	LSTM: [illegible] [illegible] NMT: [illegible]
[170]	2023	[illegible] (Encoder-decoder)	Single dialect (Algerian dialect)	BLEU	PADIC and MADAR	Transfer learning with Attention Mechanism: [illegible]
[164]	2023	[illegible] NMT Model	Multiple dialects (Levantine(LA), Maghrebi(MGR), Nile Basin Arabic, Gulf Arabic dialect and Iraqi Arabic dialect)	BLEU	PADIC, MPCA and MADAR	[illegible]
[111]	2023	Attention RNN with LSTM and Attention RNN with GRU	Single dialect (Tunisian dialect)	BLEU	Self collected dataset	Attention RNN with LSTM: [illegible] Attention RNN with GRU: [illegible]
[171]	2023	Google translate, Microsoft translator and Marian NMT model	Single dialect (Omani dialect)	BLEU	Self developed	Google: [illegible] Microsoft: [illegible] Marian NMT: [illegible] Fine tuned Marian: [illegible]
[168]	2024	Supervised encoder-decoder with attention mechanism and unsupervised transformer model	Egyptian dialect	BLEU	Self collected dataset	RNN (LSTM) with Attention (semi-supervised): [illegible] Transformer (semi-supervised): [illegible]

quality of the data, the model's size, and the integration of character-level details. Consequently, they introduce three Arabic BERT-based and two T5-based architectures. They compared their models to the latest ones on two key tasks for Arabic language processing: NLU and NLG. For NLU, they used a standard benchmark called ALUE, and for NLG, they used a part of another benchmark called ARGEN. Their models performed significantly better than existing Arabic PLMs models on both tasks.

Derouich et al. [173] performed an extensive comparison of seven different pre-trained models for sentence level translation of Arabic varieties such as, Egyptian, Emirati, Jordanian, and Palestinian vernaculars to MSA. Their proposed approach is part of the results obtained in the NADI-2023 Shared competition that exploit MADAR corpus. Specifically, their approach involved fine-tuning of the pre-trained transformer-based models including, AraT5, AraT5v2-base-1024, Sultan-ArabicT5, AraT5-MSA-Small and AraT5-MSA, AraT5-Tweet-Small and AraT5-Tweet based models. The top-performing model (AraT5v2 base-1024) attained BLEU scores of 10.02% by using the test set.

As part of the NADI-2023 shared tasks, Khered et al. [174] evaluated different T5 based networks for converting various Arabic vernaculars to MSA. Specifically, they employed three different transformer models including AraT5, multilingual T5 (mT5), and multi-task fine-tuned mT5 (mT0), which were trained jointly (with various dialects) and independently (for each dialect). Their results indicate that AraT5 model trained independently reached an outstanding BLEU score of 14.76. While, the jointly trained AraT5 model attained an outstanding BLEU point of 21.10. Their findings indicate that fine-tuning AraT5 and integrating beam search procedure results in world class translation accuracy. Similarly, [175] present their findings and results in the same shared task. The models they proposed attained a BLEU score of 11.37, earning them fourth place in the second subtask and third place in the third subtask. By employing parameter-efficient training methods, their model outperformed traditional fine-tuning approaches in the experimentation phase.

Very recently, Nacar et al. [179] introduce a technique that incorporates data augmentation techniques using generative pre-trained transformer models (GPT-3.5 and GPT-4) combined with the fine-tuning of AraT5 V2, a model specifically designed for Arabic translation. Their work is part of the OSACT 2024 Dialect to MSA Translation Shared Task. Similarly, Atwany et al. [178] reveal the findings of the same shared task. The task encompasses Gulf, Egyptian, Levantine, Iraqi, and Maghrebi dialects, providing 1001 sentences in both MSA and the dialects for fine-tuning, along with 1888 test sentences. Further experiments are conducted by fine-tuning AraT5 and No Language Left Behind (NLLB) models using the MADAR Dataset. Additionally, Fares [176] used AraT5 which achieved a BLEU score of 21.79% on the test set related to the shared task. Moreover, Alahmari [177] fine-tuned four AraT5 models for this shared task.

F. NON-ARABIC TEXT FORMALITY TRANSFER (RQ3)

Many techniques can be used for applying Arabic text formality transfer. In this subsection, we present some techniques that have not been employed in this research direction to fill more gaps through future works. We reviewed many approaches of text formality transfer that can be adapted in the domain of Arabic text formality transfer. This subsection aims to provide insight into various methodologies that can enhance the efficacy of Arabic text formality transfer.

For example, Lai et al. [182] investigated the capacity of ChatGPT as a multifaceted evaluator for the text formality transfer tasks based on existing metrics in the literature and human judgments using ZS setting. Additionally, Punkova et al. [183] adapt the advanced unsupervised-NMT model to translate dialects based on unilingual data from Croatian dialects to standardize Croatian language.

In conventional style transfer, models are limited to transferring text from one specific source style to a designated target style, facing challenges when handling texts from different domains (i.e., out-of-domain text). To address this challenge Hallinan et al. [184] present unified style transfer with an expert model that can transfer text from any given source style to numerous target styles. Their model was also developed to address the issue of limited parallel corpus. This is achieved through expert-guided decoding and a two-stage reinforcement learning approach. Their developed model surpasses the GPT-3, even though it is 226 times smaller in size.

Table 11 provides an overview of all reviewed papers on non-Arabic formality transfer according to our chosen criteria. In this table we report the best performance achieved by the presented works. The table shows the technique presented with each reviewed paper. The methodology discussed in the reviewed papers is categorized into: supervised learning (S), unsupervised learning (U), semi-supervised learning (SS), reinforcement learning (R), and LLMs. We note that Grammarly's Yahoo Answers Formality Corpus (GYAFC) [185] is the most commonly used dataset for English text.

Fig. 11 displays statistics that show the percentage breakdown of the various approaches employed in the reviewed studies. It is evident that many recent studies utilize unsupervised learning to address the issue of limited resources. It is clear also that most of reviewed papers present encoder-decoder (ED) technique as a sequence to sequence text generation. It is worth noting that the conventional approaches were not used in any of the studies reviewed for Non-Arabic text formality transfer.

It is important to highlight that many methods utilized for formality transfer in non-Arabic texts have yet to be applied to Arabic texts. This observation paves the way for researchers to explore various future opportunities in the Arabic domain. The upcoming section delves into this observation and explores various other ways in which future research could address different challenges related to formality transfer in Arabic text.

TABLE 15. Summary of MT systems used with Arabic dialect to MSA translation based on LLMs.

Ref.	Model	Dialects	Metric	Dataset	Result	Setting
[173]	AraT5-base, AraT5v2-base-1024, Sultan ArabicT5, AraT5-MSA-Small and AraT5-MSA-Base, AraT5-Tweet-Small, AraT5-Tweet-Base	Multiple dialects	BLEU	MADAR	The top performer AraT5v2-base-1024 [illegible] (with dev set) [illegible] (with test set)	Fine-tuning: Sequence length: [illegible], Learning rate (LR): 2e-5, Weight decay: 0.01
[174]	multilingual T5 (mT5), [illegible] T5, mT5 [illegible] and AraT5	Egyptian, Emirati, Jordanian, and Palestinian	BLEU	MADAR, PADIC, Dial2MSA, Arabic STS, Emi-NADI	Joint Training: mT0: 11.12, mT5: 11.13, AraT5v2: 11.36, Independent Training: mT0: 12.57, mT5: 13.27, AraT5v2: 15.14	Fine-tuning: Joint Training using various dialects, Independent Training using single dialect
[175]	AraT5	Multiple dialects	BLEU	MADAR	11.22	Fine-tuning: Learning rate: 5e-5, Weight decay: 1e-2, Sentence length: 256, 10 epochs, batch size: 2
[176]	AraT5	Egyptian, Gulf, Iraqi, Levantine, Maghrebi	BLEU	Combination of MADAR, PADIC, North Levantine, Synthetic data (GPT3)	21.79	Fine-tuning: Learning rate: 7e-5, Batch size: 32
[177]	AraT5-base, AraT5v2-base-1024, AraT5-MSA-Small, AraT5-MSA-Base	Egyptian, Gulf, Iraqi, Levantine, Maghrebi	BLEU	MADAR, PADIC, Dial2MSA, Arabic STS, SADID	23.0 (with dev set) 9.33 (with test set)	Fine-tuning: Sentence length: 128 tokens, Learning rate: 5e-5, 2 epochs, batch size: 16
[178]	GPT-3.5, AraT5, NLLB	Egyptian, Gulf, Iraqi, Levantine, Maghrebi	BLEU	MADAR	GPT-3.5: 26.81, AraT5: 19.31, NLLB: 11.56	ZS, fine-tuning
[179]	GPT-3.5, GPT-4 and AraT5 V2	Egyptian, Gulf, Iraqi, Levantine, Maghrebi	BLEU	MADAR	31.7 (with dev set) 22.6 (with test set)	Fine-tuning: Sentence length: 128 tokens, LR: 5e-5, batch size: 16

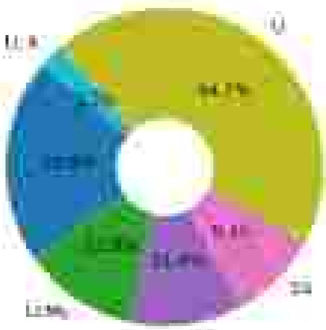

FIGURE 11. Percentage distribution of approaches used for text formality transfer.

V. DISCUSSION

Our study outlines several recent studies for text style transfer. Specifically, we considered Arabic text style transfer as a special case of MT, where informal Arabic text style (e.g., ADs) are converted (i.e., transferred) to formal Arabic text style (i.e., MSA). Thus, the study focuses on MT models between ADs and MSA, with special emphasis on the recent approaches like sequence to sequence and LLMs. Below, we consolidate the key findings, contributions, limitations, and future research gaps envisaged from these diverse approaches studied.

In this section, we address the challenges encountered in this research area and identify some research gaps that could be explored in future work. We highlight several research gaps that emerged from our analysis of the reviewed papers. The discussion primarily focuses on models that utilise sequence-to-sequence architectures and LLMs.

A. DISCUSSING THE Seq2Seq APPROACH EXPLORED FOR ARABIC TEXT FORMALITY TRANSFER

Numerous sequence-to-sequence techniques utilize transformer models to achieve exceptional translation quality. This clearly illustrates the strength of modern deep learning models, which perform MT tasks with greater accuracy compared to traditional methods like rule-based or statistical approaches.

In the context of NMT, the small range of the covered languages indicates that MT for AD is in its early stages. Certainly, all the efforts presented with the reviewed studies were focused on translating between various ADs, MSA, and English. It was noted that there is a scarcity of translations from ADs to other foreign languages.

The dialects frequently associated with this research area (as indicated in Table 5) are mainly those spoken in Algeria,

TABLE 11. Summary of reviewed papers for Non-Arabic text formality transfer.

Reference	Year	Apr.	Task	Dataset	Language	Contribution	Results (Metrics)
[illegible]	2018	S	ED	GYAFC	En	Multi-task learning	[illegible] (BLEU)
[illegible]	2019	SS	ED	GYAFC	En	Limited parallel data	[illegible] (BLEU)
[illegible]	2019	C	ED	GYAFC	En	Pseudo parallel data	[illegible]
[illegible]	2019	C	ED	GYAFC	En	Domain adaptive [illegible]	[illegible]
[illegible]	2019	R	ED	GYAFC	En	No parallel data	[illegible]
[illegible]	2019	S	[illegible]	GYAFC	En	Lexically Constrained Decoding	[illegible] (BLEU)
[illegible]	2019	S	ED	GYAFC	Fr, Pt	Limited parallel data	[illegible] (BLEU)
[illegible]	2020	SS	ED	GYAFC	En	Limited parallel data	[illegible]
[illegible]	2020	S	ED	GYAFC	En	Augmenting parallel data	[illegible]
[illegible]	2020	C	Prob	GYAFC	En	No parallel data	[illegible]
[illegible]	2020	S	ED, GPT	GYAFC	En	Limited parallel data	[illegible]
[illegible]	2020	SS	ED	GYAFC	En	Limited parallel data	[illegible] (BLEU)
[illegible]	2020	R	ED	GYAFC	En	Reinforcement learning	[illegible] (BLEU)
[illegible]	2020	S	ED	GYAFC	En	Augmenting parallel data	[illegible] (BLEU)
[illegible]	2020	C	ED	GYAFC	En	[illegible] LM	[illegible]
[illegible]	2020	C	GPT-2	GYAFC	En	Augmented generated [illegible] data	[illegible]
[illegible]	2021	C	BART	GYAFC	En	No parallel data	[illegible]
[illegible]	2021	S	ED	GYAFC	En	[illegible]	[illegible]
[illegible]	2021	S	ED	GYAFC	En	Few-shot learning	[illegible]
[illegible]	2021	C	ED	Collected	H, B, K, T, G, S, Sp	Multilingual learning	[illegible]
[illegible]	2021	LLMs	GPT	GYAFC	En	Controllable [illegible]	[illegible]
[illegible]	2021	C	KD, CL	GYAFC	En	Non-AutoRegressive decoding	[illegible]
[illegible]	2021	S	[illegible] BART	GYAFC	En	Curriculum learning	[illegible] (BLEU)
[illegible]	2021	C	Two bidirectional decoders	GYAFC	En	Collaborative learning	[illegible]
[illegible]	2021	R	[illegible]	GYAFC	En	Word disentanglement, no labelled data	[illegible]
[illegible]	2021	C	mBART	GYAFC*	F, It, Pt	No parallel data	[illegible]
[illegible]	2022	SS	ED	GYAFC	En	Limited parallel data	[illegible]
[illegible]	2022	C	BART	GYAFC	En	Roughly parallel [illegible]	[illegible]
[illegible]	2022	LLMs	[illegible] LLMs	GYAFC	En	[illegible] LM	[illegible]
[illegible]	2022	C	Prob, Pretraining classifier	GYAFC	En	Matching and editing approach	[illegible] (BLEU)
[illegible]	2022	C	BERT	GYAFC	En	Mask and [illegible] approach	[illegible]
[illegible]	2022	R	ED	N/A	N/A	N/A	N/A
[illegible]	2022	C	[illegible]	GYAFC	En	Private parallel data construction	[illegible]
[illegible]	2023	C	ED	GYAFC	En	[illegible]	[illegible]
[illegible]	2023	LLMs	[illegible]	GYAFC	En	Prompt-based editing	[illegible] (BLEU)
[illegible]	2023	C	ED	GYAFC	En	[illegible] style learning	[illegible] (PPL)
[illegible]	2023	C	NMT	Collected	Ct	First work for Croatian	[illegible] (BLEU)
[illegible]	2023	R	GPT-2 [illegible], GPT-3	GYAFC	En	[illegible]	[illegible]
[illegible]	2023	LLMs	ChatGPT	GYAFC	En	Multidimensional evaluation	[illegible]
[illegible]	2023	U, R	ED	GYAFC	En	Back Translation	[illegible]
[illegible]	2024	C	[illegible]	Raw	En	Diffusion model	[illegible]
[illegible]	2024	LLMs	[illegible]	GYAFC	En	Chain-of-thought prompting	[illegible] (BLEU)
[illegible]	2024	C	ED	GYAFC	En	[illegible] style transfer	[illegible] (PPL)

Apr.: S: Supervised learning, SS: Semi-Supervised learning, U: Unsupervised learning, R: Reinforcement learning, LLMs: Large Language Models
Task: ED: Encoder-Decoder, Prob: Probabilistic generative model, KD: Knowledge Distillation, CL: Contrastive Learning, C: [illegible]
Dataset: *: XFORMAL, collected data from Yahoo Answers, News Chat
Language: En: English, Fr: French, It: Italian, Pt: Portuguese, Ct: Croatian, H: Hindi, B: Bengali, K: Kannada, T: Telugu, G: Gujarati, S: Swahili, Sp: Spanish

Jordan, and Egypt. In contrast, dialects from regions like Saudi Arabia, Oman, Kuwait, Bahrain, and Sudan appear less frequently. Thereby, exploring performance of utilizing more Arabic dialects is a prominent research direction.

We noticed also that some of the reviewed papers focus on single dialect [159], [160], [161], leading to a shortage of models that encompass the complete range of dialects. Investigating the translation of various ADs into MSA can greatly contribute to assessing the robustness of the developed model.

From other side, most of the NMT-based AD to MSA systems reviewed in Subsection IV-E2 (see Table 9) employed single performance metric [162], [163], [164], [165]. However, relying on single evaluation measure would provide limited perspective and difficulty in comparison of the neural translation quality. Hence, for comprehensive evaluation of the NMT systems it is advisable to employ multiple evaluation metrics such as Rouge score, METEOR score, perplexity (See Subsection IV-B).

Regarding the methodology used for translating AD to MSA, RNN-based encoder-decoder is the most technique explored with seq2seq approach followed by transformer with attention mechanism. Thereby, there are more recent encoder decoder models can be explored in the future for improving performance of Arabic text formality transfer.

As a result of our analysis, we can now clearly that NMT still grapples with several weaknesses: challenges in handling extensive vocabularies, inaccuracies, or errors in translating source words, and the necessity for a substantial amount of data for training.

B. DISCUSSING THE LLMs APPROACH EXPLORED FOR ARABIC TEXT FORMALITY TRANSFER

As shown in Section IV-E3 (Table 10), there are limited studies investigating the capabilities of recent LLMs for translating ADs to MSA. In particular, the selected studies employed Arabian version of T5 model (AraT5) and GPT model (GPT-3.5 and GPT-4) to translate between ADs and MSA. Therefore, this indicates that MT between ADs and MSA based on LLMs and pre-trained transformers has just started and there is a need to explore this direction further.

In terms of the performance evaluation, all the selected studies with this category employed a single performance measure. Specifically, BLEU score is used with these selected papers. Therefore, as stated earlier depending on single performance metric could lead to bias assessment and lack of robustness in assessing the MT quality.

It is clear that there are fewer parallel corpora available for MSA and ADs compared to those available for English (see Table 7). In this regard, most researchers built their own private datasets to examine their proposed approach.

C. RESEARCH CHALLENGES AND FUTURE DIRECTIONS

This section presents a comprehensive overview of the challenges and potential future directions in Arabic text formality transfer.

1) CHALLENGES IN ARABIC TEXT FORMALITY TRANSFER

Arabic is a linguistically intricate language [9] with a diverse inflectional structure, showcased through templates and affixes, alongside various classes of attachable clitics. The Arabic language showcases intricate nuances in formality, intricately woven into its linguistic fabric and cultural landscapes. Navigating the fluidity of Arabic text poses both obstacles and openings for scholars to adjust its levels of formality.

Formality transfer encounters challenges when the initial formality of the Arabic text diverges from the desired audience or communicative setting. This discrepancy is evident across various platforms such as social media, news articles, and formal documents, where the suitable degree of formality may vary substantially.

To effectively transfer formality in Arabic text, it is crucial to go beyond just syntactic and semantic changes. Researchers must also consider cultural norms and the intended communicative purpose of the text, as ADs represent a wide range of regional dialects spoken throughout Arab countries. The emergence of social networks has substantially boosted the spread of these complicated dialects, integrating them into the fabric of daily communication. A significant amount of text generated on major social media networks such as X, Facebook, WhatsApp, and Instagram appears in these dialects.

ADs exhibit significant variation across geographic regions as (refer to Fig. 12), leading to challenges in comprehension and interpretation for those unfamiliar with a particular dialect. The language-related variations could be so substantial to the extent that in a singular nation, the same words may possess distinct interpretations. Consequently, because of the diversity among these ADs, it becomes extremely difficult to develop models that can effectively process Arabic social network platforms and other related contents.

A key approach to addressing these challenges involves capitalizing on the extensive lexicon, vast vocabulary, and well-defined structure of MSA by translating the various ADs into MSA. Whereas, the available data for Arabic text typically contains optional diacritics, spelling inconsistencies, and words written in AD is subjected to Arabization (al-Ta'rib) [224]. As a result, there is a scarcity of research concerning MT for these dialects.

The complexity of the Arabic language [225] renders the process of formality transfer very intricate, consequently diminishing the performance of such task. Navigating the formal attributes of Arabic text necessitates a nuanced grasp of its grammatical structures, lexical preferences, and stylistic norms. Arabic stands apart from numerous other languages by incorporating diverse linguistic markers and honorifics to signal formality. Consequently, numerous ADs [18] are accessible throughout the Arab world which is contributing intricate layers of complexity to the formality transfer endeavor.

FIGURE 12. Sample sentence in MSA along with their conversions into various ADs.

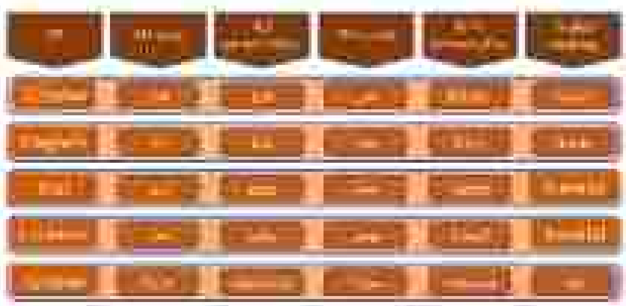

FIGURE 13. Examples of AD words and their corresponding MSA terms.

Moreover, there are various dialectal Arabic varieties at the word level. Dialectal Arabic significantly diverges from MSA phonologically, morphologically, and to a lesser extent, syntactically. Fig. 13 presents various Arabic words from different dialects alongside their MSA counterparts. It is evident from the figure that some words undergo a complete transformation when adapted to MSA. Thereby, converting AD to MSA poses numerous challenges, including the need to address these phonological shifts and morphological changes effectively. Additionally, regional idiomatic expressions and colloquialisms further complicate the conversion process, as they often lack direct MSA equivalents.

2) FUTURE WORK

As shown in Section IV-F, we reviewed very recent approaches for text formality transfer. Such approaches can be explored to improve the research prospects of Arabic text formality transfer domain. For instance, developing an innovative unified style transfer method to shift Arabic text from MSA to various ADs through out-of-domain [184] transfer presents a significant advancement, offering a promising solution to address challenges in Arabic text formality transfer.

We also observed that none of the available works in the literature studied Arabic text formality transfer using diffusion-based models. Whereas, Horvitz et al. [221] employ diffusion probabilistic models for English text formality transfer. Thus, their approach could be extended to tackle Arabic text formality transfer and check whether diffusion-based models hold an edge in respect to Arabic domain.

There remain several limitations in the field of Arabic text formality transfer that could pave the way for future research. The identified shortcomings include the following:

- Our analysis indicates that the studies examined predominantly utilize AraT5 and GPT models. Consequently, it is important to assess the effectiveness of incorporating additional LLMs, particularly those designed for Arabic text, like Arabic-llama3,[11] Jais,[12] and AceGPT.[13]
- Previous studies on Arabic text formality transfer have predominantly relied on parallel datasets for model training. Consequently, it is important to investigate the effectiveness of non-parallel datasets [168] in Arabic text formality transfer, particularly given the significant challenges associated with creating parallel datasets [226], [227].
- The earlier research on Arabic text formality transfer primarily relies on supervised learning techniques. Consequently, it is important to investigate how unsupervised [160], [228], semi-supervised, and reinforcement [215] learning methods could perform in this area of study. It is worth mentioning here that reinforcement learning can be effectively applied to the process of text formality transfer in an unsupervised learning [220]. There is also a need to explore unsupervised learning with Arabic text formality transfer by using a pseudo-parallel dataset construction [218].
- The existing research primarily relies on the BLEU score to assess the performance of Arabic text formality transfer. Therefore, it is important to consider additional evaluation metrics [229], [230], [231], [232] for a more comprehensive assessment [233]. Additionally, introducing new automatic evaluation metrics [234], [235], [236] tailored specifically for Arabic text would be highly beneficial. One approach to creating such metrics involves examining how named entities contribute to content preservation [237]. Additionally, there is a requirement to investigate how the content is

[11] https://huggingface.co/MohamedRashad/Arabic-llama3

[12] https://huggingface.co/inceptionai/Jais-Inception-7B-V1.1

[13] https://huggingface.co/FreedomIntelligence/AceGPT-13B-chat

retained [238] when applying Arabic text formality transfer. Moreover, it is important to automatically assess the coherence of the generated Arabic text.

- Given the extensive use of word embeddings in Arabic text formality transfer, it is essential to assess how different word embeddings [239], [240], [241] influence the effectiveness of this research direction.
- To improve the performance of Arabic text formality transfer, it is crucial to thoroughly examine the relationship between dialectal words and MSA words [242].
- Assessing the effectiveness of automatic word alignment tools [243] is crucial for efficiently generating large quantities of aligned parallel texts for ADs/MSA.
- Assessing how effectively different word segmentation techniques perform in the context of Arabic text formality transfer is crucial [244].
- Assessing how various Arabic parsers handle the Arabic text formality transfer is crucial [245].
- It is essential to assess how the similarities between ADs [246] influence the performance of Arabic text formality transfer.
- It is important to evaluate the impact of transliteration normalization [247] on Arabic text formality transfer.
- It is essential to develop new parallel datasets [248] for the formality transfer in Arabic text. To expand the current parallel datasets, it is important to investigate techniques for extracting parallel sentences [249].
- It is important to investigate how formality transfer relates to other style transfer tasks, like sentiment transfer [189], [190], [205], [250] and code-switching style transfer [14]. Such research may uncover innovative approaches to enhance the performance of Arabic text formality transfer.
- There is a requirement to develop additional directories [251] for Arabic machine translation.
- Improving the performance of Arabic text formality transfer requires more attention toward the Arabizi script [252].
- It is important to investigate the formality transfer in Arabic texts at the document level instead of solely focusing on the sentence level [253].

VI. CONCLUSION

This paper provides insight into research gaps within the domain of Arabic text mining, specifically focusing on formality transfer. Our research underscores the significance of exploring Arabic text formality transfer from various angles. We conducted a survey of state-of-the-art techniques in machine translation/formality transfer and devised a comparison framework to facilitate the evaluation of different methodologies. Through our analysis, we identified several gaps in the field which we have outlined in preceding sections. Additionally, our literature review reveals limitations in current Arabic formality transfer systems, many of which rely on traditional techniques with limited utilization of LLMs.

For future work, our findings suggest a shift towards evaluating newer LLMs that offer Arabic language support, such as Jais and AceGPT. In addition, addressing the scarcity of parallel datasets through non-parallel dataset analysis may offer solutions to a major research limitation. Moreover, expanding beyond supervised learning to include unsupervised, semi-supervised, and reinforcement learning methods could lead to more robust model performance. It is also valuable to employ additional evaluation metrics for a more effective assessment of content preservation and coherence, which can offer a deeper insight into the subtleties of Arabic text. Further investigation into the role of Arabic word embeddings, dialect-to-MSA relationships, and other techniques like word alignment and transliteration normalization is also crucial to advance Arabic formality transfer research.

REFERENCES

[1] A. Abdelali, H. Mubarak, S. Chowdhury, M. Hasanain, B. Mousi, S. Boughorbel, S. Abdaljalil, Y. El Kheir, D. Izham, F. Dalvi, M. Hawasly, N. Nazar, Y. Elshahawy, A. Ali, N. Durrani, N. Milic-Frayling, and F. Alam, "LAraBench: Benchmarking Arabic AI with large language models," in *Proc. 18th Conf. Eur. Chapter Assoc. Comput. Linguistics*, 2024, pp. 487–520.

[2] M. Zampieri, P. Nakov, and Y. Scherrer, "Natural language processing for similar languages, varieties, and dialects: A survey," *Natural Lang. Eng.*, vol. 26, no. 6, pp. 595–612, Nov. 2020.

[3] Z. Hu, R. K.-W. Lee, C. C. Aggarwal, and A. Zhang, "Text style transfer: A review and experimental evaluation," *ACM SIGKDD Explor. Newslett.*, vol. 24, no. 1, pp. 14–45, Jun. 2022.

[4] D. Zhang, J. Yu, P. Verma, A. Ganesan, and S. Campbell, "Improving machine translation formality control with weakly-labelled data augmentation and post editing strategies," in *Proc. 19th Int. Conf. Spoken Lang. Transl. (IWSLT)*, 2022, pp. 351–360.

[5] Y. Saoudi and M. Gammoudi, "Trends and challenges of Arabic chatbots: Literature review," *Jordanian J. Comput. Inf. Technol.*, vol. 9, no. 3, pp. 1–26, 2023.

[6] H. AlSaraie, N. AlKhamees, N. AlKhamees, M. AlHamed, N. AlThumali, S. AlZoubi, M. Ezzi, and H. AlYahya, "Arabic dialogue systems: A survey," in *Proc. 7th Int. Congr. Inf. Commun. Technol. (ICICT)*, vol. 4, London, U.K. Cham, Switzerland: Springer, 2022, pp. 153–161.

[7] D. A. Ali and N. Habash, "Botta: An Arabic dialect chatbot," in *Proc. COLING, 26th Int. Conf. Comput. Linguistics, Syst. Demonstrations*, Dec. 2016, pp. 208–212.

[8] E. Pavlick and J. Tetreault, "An empirical analysis of formality in online communication," *Trans. Assoc. Comput. Linguistics*, vol. 4, pp. 61–74, Dec. 2016.

[9] M. Alkhatib and K. Shaalan, "The key challenges for Arabic machine translation," in *Intelligent Natural Language Processing: Trends and Applications*. Cham, Switzerland: Springer, 2018, pp. 139–156.

[10] W. Salloum and N. Habash, "Dialectal Arabic to English machine translation: Pivoting through modern standard Arabic," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, Jun. 2013, pp. 348–358.

[11] E. H. Almansor and A. Al-Ani, "A hybrid neural machine translation technique for translating low resource languages," in *Proc. 14th Int. Conf. Mach. Learn. Data Mining Pattern Recognit. (MLDM)*, New York, NY, USA. Cham, Switzerland: Springer, Jan. 2018, pp. 347–356.

[12] E. Mohamed, B. Mohit, and K. Oflazer, "Transforming standard Arabic to colloquial Arabic," in *Proc. 50th Annu. Meeting Assoc. Comput. Linguistics*, Jul. 2012, pp. 176–180.

[13] I. Guellil, F. Azouaou, M. Abbas, and F. Sadat, "Arabizi transliteration of Algerian Arabic dialect into modern standard Arabic," *Social MT*, vol. 2017, pp. 1–8, May 2017.

[14] H. Xu, S. Lu, Z. Sun, C. Ma, and C. Guo, "VAE based text style transfer with pivot words enhancement learning," 2021, *arXiv:2112.03154*.

[15] A. Elnagar, S. M. Yagi, A. B. Nassif, I. Shahin, and S. A. Salloum, "Systematic literature review of dialectal Arabic: Identification and detection," IEEE Access, vol. 9, pp. 31010–31042, 2021.

[16] 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.

[17] A. Alqudsi, N. Omar, and K. Shaker, "Arabic machine translation: A survey," Artif. Intell. Rev., vol. 42, no. 4, pp. 549–572, Dec. 2014.

[18] A. M. Sabya, B. Alzahar, R. Almutlaq, and P. Elhabawy, "Textual analysis in implementation: A comprehensive review for advancing Arabic–English machine translation," in Proc. 21st Learn. Technol. Conf. (L&T), Jan. 2024, pp. 109–114.

[19] M. M. Eldesoky and T. K. Soomro, "Perspectives of Arabic machine translation," J. Eng. Sci. Technol., vol. 12, no. 9, pp. 2315–2332, 2017.

[20] S. A. Almanqrah and S. A. Alsobahi, "Challenges in rendering Arabic text to English using machine translation: A systematic literature review," IEEE Access, vol. 11, pp. 94772–94779, 2023.

[21] I. Zakraoui, M. Saleh, S. Al-Maadeed, and J. M. Alja'am, "Arabic machine translation: A survey with challenges and future directions," IEEE Access, vol. 9, pp. 161445–161468, 2021.

[22] 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.

[23] E. H. Almansor, A. Al-Ani, and F. K. Hussain, "Transferring informal text in Arabic as low resource languages: State-of-the-art and future research directions," in Proc. 13th Int. Conf. Complex, Intell., Softw. Intensive Syst. (CISIS). Cham, Switzerland: Springer, 2020, pp. 176–187.

[24] M. Qassem and M. M. Alshehri, "Can machine translate dialogue acts: Evidence from translating dialogues from English to Arabic," 3L, Lang., Linguistics, Literature, vol. 29, no. 4, pp. 65–81, Dec. 2023.

[25] E. H. Almansor, "Translating Arabic as low resource language using distribution representation and neural machine translation models," Ph.D. dissertation, School Biomed. Eng., Univ. Technol. Sydney, Sydney, NSW, Australia, 2018.

[26] W. Antoun, "Statistical machine translation of the Arabic language," Ph.D. dissertation, Comput. Sci. Lab., Univ. Le Mans, Le Mans, France, 2015.

[27] M. Kreutzer, H. Sajjad, N. Salah, A. Prupnik, F. Zemánek, and P. Pecina, "Multi-parallel corpus of north Levantine Arabic," in Proc. ArabicNLP, 2023, pp. 411–417.

[28] 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. 1–6, 2014.

[29] S. Bensalah 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.

[30] D. A. Aljehany, H. M. Al Bahanshalky, and F. A. Alsubaiee, "Arabic machine translation (ArMT) based on LSTM with attention mechanism architecture," in Proc. 20th Int. Conf. Lang. Eng. (ESOLEC), vol. 20, Oct. 2022, pp. 78–83.

[31] M. Sajjad, A. Abdelali, N. Durrani, and F. Dalvi, "AraBench: Benchmarking dialectal Arabic–English machine translation," in Proc. 28th Int. Conf. Comput. Linguistics, 2020, pp. 5094–5107.

[32] M. Elmahdy, M. Hasegawa-Johnson, and E. Mustafawi, "Development of a TV broadcasts speech recognition system for Qatari Arabic," LREC, vol. 14, pp. 3057–3061, Jan. 2014.

[33] H. Hamed, A. M. Helmy, and A. Mohammed, "Deep learning approach for translating Arabic holy Quran into Italian language," in Proc. Int. Mobile, Intell., Ubiquitous Comput. Conf. (MIUCC), May 2021, pp. 183–190.

[34] N. Bensalah, H. Ayad, A. Adib, and A. I. E. Farouk, "Arabic machine translation based on the combination of word embedding techniques," in Intelligent Systems in Big Data, Semantic Web and Machine Learning. Cham, Switzerland: Springer, 2021, pp. 247–260.

[35] A. Alrajeh, "A recipe for Arabic–English neural machine translation," 2018, arXiv:1808.06116.

[36] M. Oudah, A. Almahairi, and N. Habash, "The impact of preprocessing on Arabic–English statistical and neural machine translation," 2019, arXiv:1906.11751.

[37] N. Bensalah, H. Ayad, A. Adib, and A. I. E. Farouk, "LSTM vs. GRU for Arabic machine translation," in Proc. 12th Int. Conf. Soft Comput. Pattern Recognit. (SoCPaR). Cham, Switzerland: Springer, Jan. 2021, pp. 156–165.

[38] N. Bensalah, H. Ayad, A. Adib, and A. I. E. Farouk, "Transformer model and convolutional neural networks (CNNs) for Arabic to English machine translation," in Proc. Int. Conf. Big Data Internet Things. Cham, Switzerland: Springer, 2021, pp. 399–410.

[39] N. Bensalah, H. Ayad, A. Adib, and A. I. E. Farouk, "CRAN: An hybrid CNN-RNN attention-based model for Arabic machine translation," in Proc. Netw., Intell. Syst. Secur. (NISS). Cham, Switzerland: Springer, 2022, pp. 87–102.

[40] E. Mohamed Billah Nagoudi, A. Elmadany, and M. Abdul-Mageed, "Investigating code-mixed modern standard Arabic-Egyptian to English machine translation," 2021, arXiv:2105.13573.

[41] J. Tiedemann, "Parallel data, tools and interfaces in OPUS," LREC, vol. 2012, pp. 2214–2218, May 2012.

[42] E. M. B. Nagoudi, A. Elmadany, and M. Abdul-Mageed, "AraT5: Text-to-text transformers for Arabic language generation," 2021, arXiv:2109.12068.

[43] H. Al-Khalifa, K. Al-Khalefah, and H. Haroon, "Error analysis of pretrained language models (PLMs) in English-to-Arabic machine translation," Hum.-Centric Intell. Syst., vol. 4, no. 2, pp. 206–219, Feb. 2024.

[44] H. Alkhudur, H. Bouamor, N. Habash, and A. Zeyrek, "Benchmarking dialectal Arabic–Turkish machine translation," in Proc. Mach. Transl. Summit, vol. 1, 2023, pp. 261–271.

[45] P. Riley, T. Dozat, J. A. Botha, X. Garcia, D. Garrette, J. Riesa, O. Firat, and N. Constant, "FRMT: A benchmark for few-shot region-aware machine translation," Trans. Assoc. Comput. Linguistics, vol. 11, pp. 671–685, Jun. 2023.

[46] K. Kadaoui, S. M. Magdy, A. Waheed, M. T. I. Khondaker, A. O. El-Shangiti, E. M. B. Nagoudi, and M. Abdul-Mageed, "TARJAMAT: Evaluation of bard and ChatGPT on machine translation of ten Arabic varieties," 2023, arXiv:2308.03051.

[47] M. T. I. Khondaker, A. Waheed, E. M. B. Nagoudi, and M. Abdul-Mageed, "GPTAraEval: A comprehensive evaluation of ChatGPT on Arabic NLP," 2023, arXiv:2305.14976.

[48] Y. Moslem, R. Haque, J. D. Kelleher, and A. Way, "Adaptive machine translation with large language models," 2023, arXiv:2301.13294.

[49] L. Alkhawaja, "Unveiling the new frontier: ChatGPT-3 powered translation for Arabic–English language pairs," Theory Pract. Lang. Stud., vol. 14, no. 2, pp. 347–357, Feb. 2024.

[50] C. Lyu, J. Xu, and L. Wang, "New trends in machine translation using large language models: Case examples with ChatGPT," 2023, arXiv:2305.01181.

[51] X. Zhang, N. Rajabi, K. Duh, and P. Koehn, "Machine translation with large language models: Prompting, few-shot learning, and fine-tuning with QLoRA," in Proc. 8th Conf. Mach. Transl., 2023, pp. 468–481.

[52] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient finetuning of quantized LLMs," in Proc. Adv. Neural Inf. Process. Syst., vol. 36, 2024, pp. 1–13.

[53] Z. Alyafeai, M. S. Alshaibani, B. AlKhamissi, H. Luqman, E. Alareqi, and A. Fadel, "Taqyim: Evaluating Arabic NLP tasks using ChatGPT models," 2023, arXiv:2306.16322.

[54] M. Ziemski, M. Junczys-Dowmunt, and B. Pouliquen, "The united nations parallel corpus v1.0," in Proc. 10th Int. Conf. Lang. Resources Eval. (LREC), 2016, pp. 3530–3534.

[55] W. Jiao, W. Wang, J.-T. Huang, X. Wang, S. Shi, and Z. Tu, "Is ChatGPT a good translator? Yes with GPT-4 as the engine," 2023, arXiv:2301.08745.

[56] A. Hendy, M. Abdelrehim, A. Sharaf, V. Raunak, M. Gabr, H. Matsushita, Y. J. Kim, M. Afify, and H. H. Awadalla, "How good are GPT models at machine translation? A comprehensive evaluation," 2023, arXiv:2302.09210.

[57] Z. Tao, D. Xi, Z. Li, L. Tang, and W. Xu, "CAT-LLM: Prompting large language models with text style definition for Chinese article-style transfer," 2024, arXiv:2401.05707.

[58] M. Dare, V. Fajardo Diaz, A. Ho Zoen So, Y. Wang, and S. Zhang, "Unsupervised Mandarin-Cantonese machine translation," 2023, arXiv:2301.03971.

[59] Y. Lyu, T. Luo, J. Shi, T. C. Hollon, and H. Lee, "Fine-grained text style transfer with diffusion-based language models," 2023, arXiv:2305.19512.

[60] E. Mendes, J. F. Pimentel, L. Murta, M. Kalinowski, E. Mendes, and C. Wohlin, "On the performance of hybrid search strategies for systematic literature reviews in software engineering," Inf. Softw. Technol., vol. 123, Jul. 2020, Art. no. 106294.

[61] W. Zaghouani, "Critical survey of the freely available Arabic corpora," 2017, arXiv:1702.07835.
[62] R. Bahrain and M. Saleh, "Survey of the Arabic machine translation corpora," in Proc. Int. Conf. Modeling Implement. Complex Syst. Cham, Switzerland: Springer, Oct. 2022, pp. 205–219.
[63] J. Younes, E. Souissi, H. Achour, and A. Ferchichi, "Language resources for Maghrebi Arabic dialects' NLP: A survey," Lang. Resources Eval., vol. 54, no. 4, pp. 1079–1142, Dec. 2020.
[64] S. Harrat, K. Meftouh, and K. Smaili, "Creating parallel Arabic dialect corpus: Pitfalls to avoid," in Proc. 18th Int. Conf. Comput. Linguistics Intell. Text Process. (CICLING), Apr. 2017, pp. 1–11.
[65] R. Zbib, E. Malchiodi, J. Devlin, D. Stallard, S. Matsoukas, R. Schwartz, J. Makhoul, O. Zaidan, and C. Callison-Burch, "Machine translation of Arabic dialects," in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol., Jun. 2012, pp. 49–59.
[66] K. Smaili, M. Abbas, K. Meftouh, and S. Harrat, "Building resources for Algerian Arabic dialects," in Proc. 15th Annu. Conf. Int. Commun. Assoc. Interspeech, Sep. 2014, pp. 2123–2127.
[67] H. Bouamor, N. Habash, and K. Oflazer, "A multidialectal parallel corpus of Arabic," in Proc. LREC, May 2014, pp. 1240–1245.
[68] K. Meftouh, S. Harrat, S. Jamoussi, M. Abbas, and K. Smaili, "Machine translation experiments on PADIC: A parallel Arabic DIalect corpus," in Proc. 29th Pacific Asia Conf. Lang., Inf. Comput., Oct. 2015, pp. 26–34.
[69] K. Meftouh, S. Harrat, and K. Smaili, "PADIC: Extension and new experiments," in Proc. 7th Int. Conf. Adv. Technol. (ICAT), Apr. 2018, pp. 1–13.
[70] H. Bouamor, N. Habash, M. Salameh, W. Zaghouani, O. Rambow, D. Abdulrahim, O. Obeid, S. Khalifa, F. Eryani, A. Erdmann, and K. Oflazer, "The MADAR Arabic dialect corpus and lexicon," in Proc. 11th Int. Conf. Lang. Resources Eval. (LREC), May 2018, pp. 3387–3396.
[71] H. Mubarak, "Dial2MSA: A tweets corpus for converting dialectal Arabic to modern standard Arabic," OSACT, vol. 3, p. 49, May 2018.
[72] N. Al-Twairesh, R. Al-Matham, N. Madi, N. Almugren, A.-H. Al-Aljmi, S. Alshalan, R. Alshalan, N. Alrumayyan, S. Al-Manea, S. Bawazeer, N. Al-Mutlaq, N. Almanea, W. B. Huwaymil, D. Alqusair, R. Alotaibi, S. Al-Suwaidi, and A. Alfutamani, "SUAR: Towards building a corpus for the Saudi dialect," Proc. Comput. Sci., vol. 142, pp. 72–82, Apr. 2018.
[73] M. Abdul-Mageed, H. Alhuzali, and M. Elaraby, "You tweet what you speak: A city-level dataset of Arabic dialects," in Proc. 11th Int. Conf. Lang. Resources Eval. (LREC), May 2018, pp. 1–11.
[74] A. Abdelali, M. Attia, Y. Samih, K. Darwish, and H. Mubarak, "Diacritization of Maghrebi Arabic sub-dialects," 2018, arXiv:1810.06619.
[75] 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.
[76] S. Kchaou, R. Boujelbane, and L. H. Belguith, "Parallel resources for Tunisian Arabic dialect translation," in Proc. 5th Arabic Natural Lang. Process. Workshop, Dec. 2020, pp. 200–206.
[77] M. Al-Sulaiman, A. M. Moussa, S. Abdou, H. Elgibreen, M. Faisal, and M. Rashwan, "Semantic textual similarity for modern standard and dialectal Arabic using transfer learning," PLoS ONE, vol. 17, no. 8, Aug. 2022, Art. no. e0272991.
[78] M. Jarrar, F. A. Zaraket, T. Hammouda, D. M. Alavi, and M. Waahlisch, "Lisan: Yemeni, Iraqi, Libyan, and Sudanese Arabic dialect copora with morphological annotations," 2022, arXiv:2212.06468.
[79] A. Nayouf, T. Hammouda, M. Jarrar, F. Zaraket, and M.-B. Kurdy, "Nabra: Syrian Arabic dialects with morphological annotations," 2023, arXiv:2310.17315.
[80] A. Yamani, R. Alziyady, R. AlYami, S. Albelali, L. Albelali, J. Almulhim, A. Alsulami, M. Alfarraj, and R. Al-Zaidy, "The KIND dataset: A social collaboration approach for nuanced dialect data collection," in Proc. 18th Conf. Eur. Chapter Assoc. Comput. Linguistics, Student Res. Workshop, 2024, pp. 32–43.
[81] M. Jarrar, N. Habash, D. Akra, and N. Zalmout, "Building a corpus for Palestinian Arabic: A preliminary study," in Proc. EMNLP Workshop Arabic Natural Lang. Process. (ANLP), 2014, pp. 18–27.
[82] H. Mubarak and K. Darwish, "Using Twitter to collect a multi-dialectal corpus of Arabic," in Proc. EMNLP Workshop Arabic Natural Lang. Process. (ANLP), 2014, pp. 1–7.
[83] M. A. Sghaier and M. Zrigui, "Tunisian dialect-modern standard Arabic bilingual lexicon," in Proc. IEEE/ACS 14th Int. Conf. Comput. Syst. Appl. (AICCSA), Oct. 2017, pp. 973–979.
[84] M. Jarrar, N. Habash, F. Alrimawi, D. Akra, and N. Zalmout, "Curras: An annotated corpus for the Palestinian Arabic dialect," Lang. Resources Eval., vol. 51, no. 3, pp. 745–775, Sep. 2017.
[85] I. Guellil and A. Faical, "Bilingual lexicon for Algerian Arabic dialect treatment in social media," in Proc. Women Underrepresented Minorities Natural Lang. Process. (WiNLP), 2017, pp. 1–16.
[86] I. Alsarsour, E. Mohamed, R. Suwaileh, and T. Elsayed, "DART: A large dataset of dialectal Arabic tweets," in Proc. 11th Int. Conf. Lang. Resources Eval. (LREC), May 2018, pp. 1–8.
[87] A. Alshutayri and E. Atwell, "Creating an Arabic dialect text corpus by exploring Twitter, Facebook, and online newspapers," in Proc. OSACT, May 2018, pp. 1–8.
[88] A. Charfi, W. Zaghouani, S. H. Mehdi, and E. Mohamed, "A fine-grained annotated multi-dialectal Arabic corpus," in Proc. Natural Lang. Process. Deep Learn. World (RANLP), Oct. 2019, pp. 198–204.
[89] A. Outchakoucht and H. Es-Samaali, "Moroccan dialect–Darija–open dataset," 2021, arXiv:2103.09687.
[90] S. Dibas, C. Khairallah, N. Habash, O. F. Sadi, T. Sairafy, K. Sarabta, and A. Ardah, "Maknuune: A large open Palestinian Arabic lexicon," 2022, arXiv:2210.12985.
[91] A. C. Mazari and A. Djeffal, "Sentiment analysis of Algerian dialect using machine learning and deep learning with Word2vec," Informatica, vol. 46, no. 6, pp. 1–12, Jul. 2022.
[92] M. J. Althobaiti, "Creation of annotated country-level dialectal Arabic resources: An unsupervised approach," Natural Lang. Eng., vol. 28, no. 5, pp. 607–648, Sep. 2022.
[93] M. Alrefaie, H. Ghoneim, G. Chadhare, M. Kobeissi, and C. Ohanessian, "101 billion Arabic words dataset," 2024, arXiv:2405.01590.
[94] R. A. AlAlawani, M. Alkhatib, and K. Shaalan, "Towards Gulf Emirati dialect corpus from social media," in Proc. BUiD Doctoral Res. Conf., Multidisciplinary Stud. Cham, Switzerland: Springer, Jan. 2024, pp. 273–281.
[95] E. M. B. Nagoudi, A. Elmadany, A. El-Shangiti, and M. Abdul-Mageed, "Dolphin: A challenging and diverse benchmark for Arabic NLG," 2023, arXiv:2305.14989.
[96] A. Elmadany, E. M. B. Nagoudi, and M. Abdul-Mageed, "ORCA: A challenging benchmark for Arabic language understanding," 2022, arXiv:2212.10758.
[97] H. Al-Khalifa, K. Darwish, H. Mubarak, M. Ali, and T. Elsayed, "The 6th workshop on open-source Arabic corpora and processing tools (OSACT) with shared tasks on Arabic LLMs hallucination and dialect to MSA machine translation," in Proc. 6th Workshop Open-Source Arabic Corpora Process. Tools (OSACT) Shared Tasks Arabic LLMs Hallucination Dialect MSA Mach. Transl. (LREC-COLING), 2024, pp. 1–10.
[98] M. Abdul-Mageed, A. Elmadany, C. Zhang, E. M. B. Nagoudi, H. Bouamor, and N. Habash, "NADI 2023: The fourth nuanced Arabic dialect identification shared task," 2023, arXiv:2310.16117.
[99] M. Abdul-Mageed, C. Zhang, A. Elmadany, H. Bouamor, and N. Habash, "NADI 2022: The third nuanced Arabic dialect identification shared task," 2022, arXiv:2210.09582.
[100] M. Abdul-Mageed, C. Zhang, H. Bouamor, and N. Habash, "NADI 2020: The first nuanced Arabic dialect identification shared task," 2020, arXiv:2010.11334.
[101] N. Habash and O. Rambow, "MAGEAD: A morphological analyzer and generator for the Arabic dialects," in Proc. 21st Int. Conf. Comput. Linguistics 44th Annu. Meeting (ACL), 2006, pp. 681–688.
[102] H. Abo Bakr, K. Shaalan, and I. Ziedan, "A hybrid approach for converting written Egyptian colloquial dialect into diacritized Arabic," in Proc. 6th Int. Conf. Informat. Syst., Jan. 2008, pp. 1–14.
[103] M. Diab, N. Habash, O. Rambow, M. Altantawy, and Y. Benajiba, "COLABA: Arabic dialect annotation and processing," in Proc. LREC Workshop Semitic Lang. Process., 2010, pp. 66–74.
[104] H. Sawaf, "Arabic dialect handling in hybrid machine translation," in Proc. 9th Conf. Assoc. Mach. Transl. Americas: Res. Papers, Jan. 2010, pp. 1–8.
[105] W. Salloum and N. Habash, "Dialectal to standard Arabic paraphrasing to improve Arabic–English statistical machine translation," in Proc. 1st Workshop Algorithms Resources Modelling Dialects Lang. Varieties, Jul. 2011, pp. 10–21.
[106] O. F. Zaidan and C. Callison-Burch, "The Arabic online commentary dataset: An annotated dataset of informal Arabic with high dialectal content," in Proc. 49th Annu. Meeting Assoc. Comput. Linguistics, Hum. Lang. Technol., Jun. 2011, pp. 37–41.

[107] W. Salloum, H. Elfardy, L. Alamir-Salloum, N. Habash, and M. Diab, "Sentence level dialect identification for machine translation system selection," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics*, 2014, pp. 772–778.

[108] F. Sadat, "Multi-dialect machine translation (MuDMat)," in *Proc. 18th Annu. Conf. Eur. Assoc. Mach. Transl.*, May 2015, p. 226.

[109] B. Torjmen, N. G. Hammouda, and K. Haddar, "A Novel Tunisian dialect translation," in *Proc. 13th Int. Conf. Formalizing Natural Lang. Word, Natural Lang. Process. Appl.*, Hammamet, Tunisia. Cham, Switzerland: Springer, 2019, pp. 123–134.

[110] M. Lichouri, R. Djeradi, A. Djeradi, and M. Abbas, "Towards a portable SLU system applied to MSA and low-resourced Algerian dialects," in *Proc. Int. Conf. Adv. Mach. Learn. Technol. Appl. (AMLTA)*. Cham, Switzerland: Springer, 2020, pp. 576–585.

[111] T. Alshareef and M. A. Siddiqui, "A seq2seq neural network based conversational agent for Gulf Arabic dialect," in *Proc. 21st Int. Arab Conf. Inf. Technol. (ACIT)*, Nov. 2020, pp. 1–7.

[112] O. Obeid, N. Zalmout, S. Khalifa, D. Taji, M. Oudah, B. Alhafni, G. Inoue, F. Eryani, A. Erdmann, and N. Habash, "CAMeL tools: An open source Python toolkit for Arabic natural language processing," in *Proc. 12th Lang. Resources Eval. Conf.*, May 2020, pp. 7022–7032.

[113] A. Gargett and T. T.-C. Leung, "Building the Emirati Arabic FrameNet," in *Proc. Int. FrameNet Workshop, Towards Global, Multilingual FrameNets*, May 2020, pp. 70–78.

[114] Y. M. N. Sultan, M. S. M. Hassan, R. Elhalfi, and A. Omar, "An evaluation of the accuracy of the machine translation systems of social media language," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 7, pp. 1–06, 2021.

[115] J. Chen, O. Aasejo, and W. Zhao, "An integrated participatory platform for human evaluation of machine translation," in *Proc. 12th ACM/IEEE-CS Joint Conf. Digit. Libraries*, Jun. 2012, pp. 421–422.

[116] J. Zakraoui, M. Saleh, S. Al-Maadeed, and J. M. AlJa'am, "Evaluation of Arabic to English machine translation systems," in *Proc. 11th Int. Conf. Inf. Commun. Syst. (ICICS)*, Apr. 2020, pp. 185–190.

[117] A. Lavie and M. J. Denkowski, "The meteor metric for automatic evaluation of machine translation," *Mach. Transl.*, vol. 23, nos. 2–3, pp. 105–115, Sep. 2009.

[118] 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*, 2002, pp. 311–318.

[119] R. Shu, H. Nakayama, and K. Cho, "Generating diverse translations with sentence codes," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 1823–1827.

[120] F. Xiao, L. Pang, Y. Lan, Y. Wang, H. Shen, and X. Cheng, "Transductive learning for unsupervised text style transfer," 2021, *arXiv:2109.07812*.

[121] M. Snover, B. J. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A study of translation edit rate with targeted human annotation," in *Proc. 7th Conf. Assoc. Mach. Transl. Americas, Tech. Papers*, Aug. 2006, pp. 223–231.

[122] M. G. Snover, N. Madnani, B. Dorr, and R. Schwartz, "TER-plus: Paraphrase, semantic, and alignment enhancements to translation edit rate," *Mach. Transl.*, vol. 23, nos. 2–3, pp. 117–127, Sep. 2009.

[123] R. Rei, C. Stewart, A. C. Farinha, and A. Lavie, "COMET: A neural framework for MT evaluation," 2020, *arXiv:2009.09025*.

[124] M. Popović, "ChrF: Character n-gram F-score for automatic MT evaluation," in *Proc. 10th Workshop Stat. Mach. Transl.*, 2015, pp. 392–395.

[125] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating text generation with BERT," 2019, *arXiv:1904.09675*.

[126] L. Chin-Yew, "Rouge: A package for automatic evaluation of summaries," in *Proc. Workshop Text Summarization Branches Out*, 2004, pp. 1–20.

[127] D. Colla, M. Delsanto, M. Agosto, B. Vitiello, and D. P. Radicioni, "Semantic coherence markers: The contribution of perplexity metrics," *Artif. Intell. Med.*, vol. 134, Dec. 2022, Art. no. 102393.

[128] C. Napoles, K. Sakaguchi, M. Post, and J. Tetreault, "Ground truth for grammatical error correction metrics," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics 7th Int. Joint Conf. Natural Lang. Process.*, 2015, pp. 588–593.

[129] Y. Cheng, Z. Gan, Y. Zhang, O. Elachqar, D. Li, and J. Liu, "Contextual text style transfer," 2020, *arXiv:2005.00136*.

[130] D. Chen and W. B. Dolan, "Collecting highly parallel data for paraphrase evaluation," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics, Hum. Lang. Technol.*, Jun. 2011, pp. 190–200.

[131] Z. Fu, X. Tan, N. Peng, D. Zhao, and R. Yan, "Style transfer in text: Exploration and evaluation," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2018, vol. 32, no. 1, pp. 1–18.

[132] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," 2019, *arXiv:1908.10084*.

[133] V. John, L. Mou, H. Bahuleyan, and O. Vechtomova, "Disentangled representation learning for non-parallel text style transfer," 2018, *arXiv:1808.04339*.

[134] W. Yuan, G. Neubig, and P. Liu, "BARTScore: Evaluating generated text as text generation," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2021, pp. 27263–27277.

[135] T. Sellam, D. Das, and A. P. Parikh, "BLEURT: Learning robust metrics for text generation," 2020, *arXiv:2004.04696*.

[136] A. Warstadt, A. Singh, and S. R. Bowman, "Neural network acceptability judgments," *Trans. Assoc. Comput. Linguistics*, vol. 7, pp. 625–641, Nov. 2019.

[137] R. Mir, B. Felbo, N. Obradovich, and I. Rahwan, "Evaluating style transfer for text," 2019, *arXiv:1904.02295*.

[138] J. Wieting, T. Berg-Kirkpatrick, K. Gimpel, and G. Neubig, "Beyond BLEU: Training neural machine translation with semantic similarity," 2019, *arXiv:1909.06694*.

[139] E. Agirre, C. Banea, D. Cer, M. Diab, A. Gonzalez-Agirre, R. Mihalcea, G. Rigau, and J. Wiebe, "SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation," in *Proc. 10th Int. Workshop Semantic Eval. (SemEval)*, San Diego, CA, USA, 2016, pp. 497–511.

[140] D. Jin, Z. Jin, Z. Hu, O. Vechtomova, and R. Mihalcea, "Deep learning for text style transfer: A survey," *Comput. Linguistics*, vol. 48, no. 1, pp. 155–205, Apr. 2022.

[141] W. Salloum, *Machine Translation of Arabic Dialects*. New York, NY, USA: Columbia Univ. Press, 2018.

[142] M. Lichouri and M. Abbas, "Machine translation for zero and low-resourced dialects using a new extended version of the dialectal parallel corpus (PADIC v2.0)," in *Proc. 4th Int. Conf. Natural Lang. Speech Process. (ICNLSP)*, 2021, pp. 55–58.

[143] A. Pandey, K. K. Yogi, and R. Sharma, "A comparison of machine translation methods for natural language processing and their challenges," in *Proc. Int. Conf. Innov. Comput. Intell. Comput. Vol.* Cham, Switzerland: Springer, 2022, pp. 473–487.

[144] E. H. Almansor and A. Al-Ani, "Translating dialectal Arabic as low resource language using word embedding," in *Proc. Int. Conf. Recent Adv. Natural Lang. Process. (RANLP)*, Nov. 2017, pp. 52–57.

[145] F. Sadat, F. Mallek, M. Boudabous, R. Sellami, and A. Farzindar, "Collaboratively constructed linguistic resources for language variants and their exploitation in NLP application—The case of Tunisian Arabic and the social media," in *Proc. Workshop Lexical Grammatical Resources Lang. Process.*, 2014, pp. 102–110.

[146] S. Hamada and R. M. Marzouk, "Developing a transfer-based system for Arabic dialects translation," in *Intelligent Natural Language Processing: Trends and Applications*, 2018, pp. 121–138.

[147] A. Hamdi, R. Boujelbane, N. Habash, and A. Nasr, "The effects of factorizing root and pattern mapping in bidirectional Tunisian-standard Arabic machine translation," in *Proc. Mach. Transl. Summit*, 2013, pp. 1–21.

[148] M. A. Sghaier and M. Zrigui, "Rule-based machine translation from Tunisian dialect to modern standard Arabic," *Proc. Comput. Sci.*, vol. 176, pp. 310–319, Mar. 2020.

[149] W. Salloum and N. Habash, "Elissa: A dialectal to standard Arabic machine translation system," in *Proc. COLING, Demonstration Papers*, Dec. 2012, pp. 385–392.

[150] G. Al-Gaphari and M. Al-Yadoumi, "A method to convert Sana'ani accent to modern standard Arabic," *Int. J. Inf. Sci. Manage.*, vol. 8, no. 1, pp. 39–49, 2010.

[151] S. Alsamdi and F. Ahmed, "Sana'ani dialect to modern standard Arabic: Rule-based direct machine translation," in *Proc. Int. Conf. Artif. Intell. (ICAI)*, 2011, pp. 1–4.

[152] S. Bahti, O. Amraoui, N. E. Mouhib, R. Kirchoum, and Y. Chihab, "Moroccan Arabizi-to-Arabic conversion using rule-based transliteration and weighted Levenshtein algorithm," *Sci. Afr.*, vol. 23, Mar. 2024, Art. no. e02073.

[153] S. Jeblee, W. Feely, H. Bouamor, A. Lavie, N. Habash, and K. Oflazer, "Domain and dialect adaptation for machine translation into Egyptian Arabic," in *Proc. EMNLP Workshop Arabic Natural Lang. Process. (ANLP)*, 2014, pp. 196–206.
[154] H. Sajjad, K. Darwish, and Y. Belinkov, "Translating dialectal Arabic to English," in *Proc. 51st Annu. Meeting Assoc. Comput. Linguistics*, Jun. 2013, pp. 1–6.
[155] P. Nakov and H. T. Ng, "Improving statistical machine translation for a resource-poor language using related resource-rich languages," *J. Artif. Intell. Res.*, vol. 44, pp. 179–222, May 2012.
[156] A. Alnassan, "Automatic standardization of Arabic dialects for machine translation," 2023, *arXiv:2301.03447*.
[157] R. Tachicart and K. Bouzoubaa, "A hybrid approach to translate Moroccan Arabic dialect," in *Proc. 9th Int. Conf. Intell. Syst., Theories Appl. (SITA)*, May 2014, pp. 1–5.
[158] H. Trost, "A hybrid approach to statistical machine translation between standard and dialectal varieties," in *Proc. 6th Lang. Technol. Conf. Hum. Lang. Technol. Challenges Comput. Sci. Linguistics (LTC)*, vol. 9561. Poznań, Poland. Cham, Switzerland: Springer, Dec. 2013, p. 341.
[159] I. Guellil, F. Azouaou, and M. Abbas, "Neural vs statistical translation of Algerian Arabic dialect written with Arabizi and Arabic letter," in *Proc. 31st Pacific Asia Conf. Lang., Inf. Comput. Paclic*, vol. 31, 2017, pp. 1–27.
[160] 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.
[161] R. Al-Ibrahim and R. M. Duwairi, "Neural machine translation from Jordanian dialect to modern standard Arabic," in *Proc. 11th Int. Conf. Inf. Commun. Syst. (ICICS)*, Apr. 2020, pp. 173–178.
[162] L. H. Baniata, S. Park, and S.-B. Park, "A neural machine translation model for Arabic dialects that utilizes multitask learning (MTL)," *Comput. Intell. Neurosci.*, vol. 2018, pp. 1–10, Dec. 2018.
[163] L. H. Baniata, S. Park, and S.-B. Park, "A multitask-based neural machine translation model with part-of-speech tags integration for Arabic dialects," *Appl. Sci.*, vol. 8, no. 12, p. 2502, Dec. 2018.
[164] L. H. Baniata, S. Kang, and I. K. E. Ampomah, "A reverse positional encoding multi-head attention-based neural machine translation model for Arabic dialects," *Mathematics*, vol. 10, no. 19, p. 3666, Oct. 2022.
[165] L. H. Baniata, I. K. E. Ampomah, and S. Park, "A transformer-based neural machine translation model for Arabic dialects that utilizes subword units," *Sensors*, vol. 21, no. 19, p. 6509, Sep. 2021.
[166] A. Fsih, S. Kchaou, and R. Boujelbane, "Neural machine translation of low resource languages: Application to transcriptions of Tunisian dialect," in *Proc. Int. Conf. Intell. Syst. Pattern Recognit.* Cham, Switzerland: Springer, Jan. 2022, pp. 234–247.
[167] S. Kchaou, R. Boujelbane, and L. Hadrich, "Hybrid pipeline for building Arabic Tunisian dialect-standard Arabic neural machine translation model from scratch," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 22, no. 3, pp. 1–21, Mar. 2023.
[168] M. A. Faheem, K. T. Wassif, H. Bayomi, and S. M. Abdou, "Improving neural machine translation for low resource languages through non-parallel corpora: A case study of Egyptian dialect to modern standard Arabic translation," *Sci. Rep.*, vol. 14, no. 1, p. 2265, Jan. 2024.
[169] A. Slim, A. Melouah, Y. Faghihi, and K. Sahib, "Algerian dialect translation applied on COVID-19 social media comments," in *Artificial Intelligence and Renewables Towards an Energy Transition*. Cham, Switzerland: Springer, 2021, pp. 716–726.
[170] A. Slim, A. Melouah, U. Faghihi, and K. Sahib, "Improving neural machine translation for low resource Algerian dialect by transductive transfer learning strategy," *Arabian J. Sci. Eng.*, vol. 47, no. 8, pp. 10411–10418, Aug. 2022.
[171] K. Al-Kharusi and A. AAlAbdulsalam, "Machine translation of Omani Arabic dialect from social media," in *Proc. ArabicNLP*, 2023, pp. 302–309.
[172] Y. Moukafih, N. Sbihi, M. Ghogho, and K. Smaili, "Improving machine translation of Arabic dialects through multi-task learning," in *Proc. Int. Conf. Italian Assoc. Artif. Intell.* Cham, Switzerland: Springer, 2021, pp. 580–590.
[173] W. Derouich, S. Kchaou, and R. Boujelbane, "ANLP-RG at NADI 2023 shared task: Machine translation of Arabic dialects: A comparative study of transformer models," in *Proc. ArabicNLP*, 2023, pp. 683–689.
[174] A. Khered, I. Abdelhalim, N. Abdelhalim, A. Soliman, and R. Batista-Navarro, "UniManc at NADI 2023 shared task: A comparison of various T5-based models for translating Arabic dialectical text to modern standard Arabic," in *Proc. ArabicNLP*, 2023, pp. 658–664.
[175] R. Abdel-Salam, "Rematchka at NADI 2023 shared task: Parameter efficient tuning for dialect identification and dialect machine translation," in *Proc. ArabicNLP*, 2023, pp. 652–657.
[176] M. Fares, "AraT5-MSAizer: Translating dialectal Arabic to MSA," in *Proc. 6th Workshop Open-Source Arabic Corpora Process. Tools (OSACT) Shared Tasks Arabic LLMs Hallucination Dialect MSA Mach. Transl. (LREC-COLING)*, 2024, pp. 124–129.
[177] S. S. Alahmari, "Sirius_translators at OSACT6 2024 shared task: Fin-tuning AraT5 models for translating Arabic dialectal text to modern standard Arabic," in *Proc. 6th Workshop Open-Source Arabic Corpora Process. Tools (OSACT) Shared Tasks Arabic LLMs Hallucination Dialect MSA Mach. Transl. (LREC-COLING)*, 2024, pp. 117–123.
[178] H. Atwany, N. Rabih, I. Mohammed, A. Waheed, and B. Raj, "OSACT 2024 task 2: Arabic dialect to MSA translation," in *Proc. 6th Workshop Open-Source Arabic Corpora Process. Tools (OSACT) Shared Tasks Arabic LLMs Hallucination Dialect MSA Mach. Transl. (LREC-COLING)*, 2024, pp. 98–103.
[179] O. Nacar, A. Alharbi, S. Sibaee, S. Ahmed, L. Ghouti, and A. Koubaa, "ASOS at OSACT6 shared task: Investigation of data augmentation in Arabic dialect-MSA translation," in *Proc. 6th Workshop Open-Source Arabic Corpora Process. Tools (OSACT) Shared Tasks Arabic LLMs Hallucination Dialect MSA Mach. Transl. (LREC-COLING)*, 2024, pp. 104–111.
[180] J. Zhao, Z. Zhang, L. Gao, Q. Zhang, T. Gui, and X. Huang, "LLaMA beyond English: An empirical study on language capability transfer," 2024, *arXiv:2401.01055*.
[181] A. Ghaddar, Y. Wu, S. Bagga, A. Rashid, K. Bibi, M. Rezagholizadeh, C. Xing, Y. Wang, X. Duan, Z. Wang, B. Huai, X. Jiang, Q. Liu, and P. Langlais, "Revisiting pre-trained language models and their evaluation for Arabic natural language processing," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2022, pp. 3135–3151.
[182] H. Lai, A. Toral, and M. Nissim, "Multidimensional evaluation for text style transfer using ChatGPT," 2023, *arXiv:2304.13462*.
[183] B. Penkova, M. Mitreska, K. Ristov, K. Mishev, and M. Simjanoska, "Learning translation model to translate croatian dialects to modern croatian language," in *Proc. 46th MIPRO ICT Electron. Conv. (MIPRO)*, May 2023, pp. 1083–1088.
[184] S. Hallinan, F. Brahman, X. Lu, J. Jung, S. Welleck, and Y. Choi, "STEER: Unified style transfer with expert reinforcement," 2023, *arXiv:2311.07167*.
[185] S. Rao and J. Tetreault, "Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer," 2018, *arXiv:1803.06535*.
[186] X. Niu, S. Rao, and M. Carpuat, "Multi-task neural models for translating between styles within and across languages," 2018, *arXiv:1806.04357*.
[187] R. Xu, T. Ge, and F. Wei, "Formality style transfer with hybrid textual annotations," 2019, *arXiv:1903.06353*.
[188] Z. Jin, D. Jin, J. Mueller, N. Matthews, and E. Santus, "IMaT: Unsupervised text attribute transfer via iterative matching and translation," 2019, *arXiv:1901.11333*.
[189] D. Li, Y. Zhang, Z. Gan, Y. Cheng, C. Brockett, M.-T. Sun, and B. Dolan, "Domain adaptive text style transfer," 2019, *arXiv:1908.09395*.
[190] H. Gong, S. Bhat, L. Wu, J. Xiong, and W.-M. Hwu, "Reinforcement learning based text style transfer without parallel training corpus," 2019, *arXiv:1903.10671*.
[191] T. Kajiwara, "Negative lexically constrained decoding for paraphrase generation," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 6047–6052.
[192] X. Niu, "Formality style transfer within and across languages with limited supervision," Ph.D. dissertation, Dept. Computer Science, Univ. Maryland, College Park, MD, USA, 2019.
[193] X. Niu and M. Carpuat, "Controlling neural machine translation formality with synthetic supervision," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2020, vol. 34, no. 5, pp. 8568–8575.
[194] J. He, X. Wang, G. Neubig, and T. Berg-Kirkpatrick, "A probabilistic formulation of unsupervised text style transfer," 2020, *arXiv:2002.03912*.
[195] Y. Wang, Y. Wu, L. Mou, Z. Li, and W. Chao, "Formality style transfer with shared latent space," in *Proc. 28th Int. Conf. Comput. Linguistics*, 2020, pp. 2236–2249.

[196] K. Chawla and D. Yang, "Semi-supervised formality style transfer using language model discriminator and mutual information maximization," 2020, arXiv:2010.05090.

[197] A. Sancheti, K. Krishna, B. Srinivasan, and A. Natarajan, "Reinforced rewards framework for text style transfer," in Proc. 42nd Eur. Conf. Adv. Inf. Retr. (ECIR), Lisbon, Portugal. Cham, Switzerland: Springer, Jan. 2020, pp. 545–560.

[198] Y. Zhang, T. Ge, and X. Sun, "Parallel data augmentation for formality style transfer," 2020, arXiv:2005.07522.

[199] N. Goyal, B. V. Srinivasan, A. Natarajan, and A. Sancheti, "Multi-style transfer with discriminative feedback on disjoint corpus," 2020, arXiv:2010.11578.

[200] K. Krishna, J. Wieting, and M. Iyyer, "Reformulating unsupervised style transfer as paraphrase generation," 2020, arXiv:2010.05700.

[201] H. Lai, A. Toral, and M. Nissim, "Generic resources are what you need: Style transfer tasks without task-specific parallel training data," 2021, arXiv:2109.04543.

[202] Z. Hu, R. Ka-Wei Lee, and C. C. Aggarwal, "Syntax matters! Syntax-controlled in text style transfer," 2021, arXiv:2108.05869.

[203] K. Krishna, D. Nathani, X. Garcia, B. Samanta, and P. Talukdar, "Few-shot controllable style transfer for low-resource multilingual settings," 2021, arXiv:2110.07385.

[204] S. Kumar, E. Malmi, A. Severyn, and Y. Tsvetkov, "Controlled text generation as continuous optimization with multiple constraints," in Proc. Adv. Neural Inf. Process. Syst., vol. 34, Jan. 2021, pp. 14542–14554.

[205] Y. Ma and Q. Li, "Exploring non-autoregressive text style transfer," in Proc. Conf. Empirical Methods Natural Lang. Process., 2021, pp. 9267–9278.

[206] S. Kadotani, T. Kajiwara, Y. Arase, and M. Onizuka, "Edit distance based curriculum learning for paraphrase generation," in Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process., Student Res. Workshop, 2021, pp. 229–234.

[207] Y. Ma, Y. Chen, X. Mao, and Q. Li, "Collaborative learning of bidirectional decoders for unsupervised text style transfer," in Proc. Conf. Empirical Methods Natural Lang. Process., 2021, pp. 9250–9266.

[208] W. Qian, J. Yang, and S. Hu, "Discovering the style information in texts via a reinforced decision process," in Proc. Int. Joint Conf. Neural Netw. (IJCNN), Jul. 2022, pp. 1–8.

[209] H. Lai, A. Toral, and M. Nissim, "Multilingual pre-training with language and task adaptation for multilingual text style transfer," 2022, arXiv:2203.08552.

[210] A. Liu, A. Wang, and N. Okazaki, "Semi-supervised formality style transfer with consistency training," 2022, arXiv:2203.13620.

[211] R. Liu, C. Gao, C. Jia, G. Xu, and S. Vosoughi, "Non-parallel text style transfer with self-parallel supervision," 2022, arXiv:2204.08123.

[212] F. Mireshghallah, K. Goyal, and T. Berg-Kirkpatrick, "Mix and match: Learning-free controllable text generation using energy language models," 2022, arXiv:2203.13299.

[213] S. Yang, "Tagging without rewriting: A probabilistic model for unpaired sentiment and style transfer," in Proc. 12th Workshop Comput. Approaches Subjectivity, Sentiment Social Media Anal., 2022, pp. 293–303.

[214] S. Yang, "Mask and regenerate: A classifier-based approach for unpaired sentiment transformation of reviews for electronic commerce websites," in Proc. 10th Int. Workshop Natural Lang. Process. Social Media, 2022, pp. 1–10.

[215] K. G. Rao, B. Shrestha, M. Vinnidhya, and A. V. V. Kumar, "Scrutiny of natural language processing using reinforcement learning," NeuroQuantology, vol. 20, no. 17, p. 1360, 2022.

[216] X. Shen, W. Chen, and F. Xu, "Unsupervised sentiment and style transfer from massive texts," in Proc. 4th Int. Conf. Electron. Inf. Technol. Comput. Eng., Oct. 2022, pp. 206–213.

[217] A. Scalercio and A. Paes, "Masked transformer through knowledge distillation for unsupervised text style transfer," Natural Lang. Eng., vol. 30, no. 5, pp. 973–1008, Sep. 2024.

[218] G. Luo, Y. Tong Han, L. Mou, and M. Firdaus, "Prompt-based editing for text style transfer," 2023, arXiv:2301.11997.

[219] L. Wu, P. Liu, Y. Yuan, S. Liu, and Y. Zhang, "Context-aware style learning and content recovery networks for neural style transfer," Inf. Process. Manage., vol. 60, no. 3, May 2023, Art. no. 103265.

[220] D. Bandyopadhyay and A. Ekbal, "Unsupervised text style transfer through differentiable back translation and rewards," in Proc. Pacific-Asia Conf. Knowl. Discovery Data Mining. Cham, Switzerland: Springer, Jan. 2023, pp. 210–221.

[221] Z. Horvitz, A. Patel, C. Callison-Burch, Z. Yu, and K. McKeown, "ParaGuide: Guided diffusion paraphrasers for plug-and-play textual style transfer," in Proc. AAAI Conf. Artif. Intell., Mar. 2024, vol. 38, no. 16, pp. 18216–18224.

[222] C. Zhang, H. Cai, Yuezhong Li, Y. Wu, L. Hou, and M. Abdul-Mageed, "Distilling text style transfer with self-explanation from LLMs," 2024, arXiv:2403.01106.

[223] Y. Gao, Q. Liu, Y. Yang, and K. Wang, "Latent representation discrimination for unsupervised text style generation," Inf. Process. Manage., vol. 61, no. 3, May 2024, Art. no. 103643.

[224] M. R. Khalid, "Al-Ta'rib: Pro and con of foreign words arabization," in Proc. 2nd Int. Conf. Culture Lang. Southeast Asia (ICCLAS), Amsterdam, The Netherlands: Atlantis Press, 2019, pp. 140–143.

[225] K. Shaalan, S. Siddiqui, M. Alkhatib, and A. A. Monem, "Challenges in Arabic natural language processing," in Computational Linguistics, Speech And Image Processing For Arabic Language. Singapore: World Scientific, 2019, pp. 59–83.

[226] O. Nabil, S. Ouaghzera, N. Khili, and B. Ouhbi, "Challenges and progress in constructing Arabic dialect corpora and linguistic tools: A focus on Moroccan and Tunisian dialects," in Proc. 7th IEEE Congr. Inf. Sci. Technol. (CiSt), Dec. 2023, pp. 293–298.

[227] M. Omari and N. Benhamou, "Challenges in building corpora for Algerian Arabic from CMC content," J. Soc. Hum. Sci., vol. 22, no. 4, pp. 594–617, 2022.

[228] W. Salloum and N. Habash, "Unsupervised Arabic dialect segmentation for machine translation," Natural Lang. Eng., vol. 28, no. 2, pp. 223–248, Mar. 2022.

[229] H. Yu, W. Xu, S. Lin, and Q. Liu, "Machine translation evaluation metric based on dependency parsing model," ACM Trans. Asian Low-Resource Lang. Inf. Process., vol. 18, no. 4, pp. 1–15, Dec. 2019.

[230] M. Popović, "Class error rates for evaluation of machine translation output," in Proc. 7th Workshop Stat. Mach. Transl., Jun. 2012, pp. 71–75.

[231] E. Briakou, S. Agrawal, J. Tetreault, and M. Carpuat, "Evaluating the evaluation metrics for style transfer: A case study in multilingual formality transfer," 2021, arXiv:2110.10668.

[232] H. Lai, J. Mao, A. Toral, and M. Nissim, "Human judgement as a compass to navigate automatic metrics for formality transfer," 2022, arXiv:2204.07549.

[233] S. Gehrmann, E. Clark, and T. Sellam, "Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text," 2022, arXiv:2202.06935.

[234] M. Li and M. Wang, "Optimizing automatic evaluation of machine translation with the ListMLE approach," ACM Trans. Asian Low-Resource Lang. Inf. Process., vol. 18, no. 1, pp. 1–18, Nov. 2018.

[235] N. Bach, F. Huang, and Y. Al-Onaizan, "Goodness: A method for measuring machine translation confidence," in Proc. 49th Annu. Meeting Assoc. Comput. Linguistics, Hum. Lang. Technol., Jun. 2011, pp. 211–219.

[236] P. Li, C. Chen, W. Zheng, Y. Deng, F. Ye, and Z. Zheng, "STD: An automatic evaluation metric for machine translation based on word embeddings," IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 27, no. 10, pp. 1497–1506, Oct. 2019.

[237] N. Babakov, D. Dale, V. Logacheva, I. Krotova, and A. Panchenko, "Studying the role of named entities for content preservation in text style transfer," in Proc. Int. Conf. Appl. Natural Lang. Inf. Syst. Cham, Switzerland: Springer, Jun. 2022, pp. 437–448.

[238] N. Babakov, D. Dale, I. Gusev, I. Krotova, and A. Panchenko, "Don't lose the message while paraphrasing: A study on content preserving style transfer," in Proc. Int. Conf. Appl. Natural Lang. Inf. Syst. Cham, Switzerland: Springer, 2023, pp. 47–61.

[239] N. Bensalah, H. Ayad, A. Adib, and A. I. E. Farouk, "A comparative study of different dimensionality reduction techniques for Arabic machine translation," ACM Trans. Asian Low-Resource Lang. Inf. Process., vol. 22, no. 12, pp. 1–17, Dec. 2023.

[240] D. S. Asudani, N. K. Nagwani, and P. Singh, "Impact of word embedding models on text analytics in deep learning environment: A review," Artif. Intell. Rev., vol. 56, no. 9, pp. 10345–10425, Sep. 2023.

[241] S. Jafaritazehjani, G. Lecorvé, D. Lolive, and J. D. Kelleher, "Local or global: The variation in the encoding of style across sentiment and formality," in Proc. Int. Conf. Artif. Neural Netw. Cham, Switzerland: Springer, Jan. 2023, pp. 492–504.
[242] S. Harrat, K. Meftouh, M. Abbas, S. Jamoussi, M. Saad, and K. Smaïli, "Cross-dialectal Arabic processing," in Proc. 16th Int. Conf. Comput. Linguistics Intell. Text Process. (CICLing), Cairo, Egypt. Cham, Switzerland: Springer, Jan. 2015, pp. 620–632.
[243] S. Grimes, K. Peterson, and X. Li, "Automatic word alignment tools to scale production of manually aligned parallel texts," in Proc. LREC, May 2012, pp. 2196–2198.
[244] M. Paul, A. Finch, and E. Sumita, "Word segmentation for dialect translation," in Proc. Int. Conf. Intell. Text Process. Comput. Linguistics. Cham, Switzerland: Springer, Jan. 2011, pp. 55–67.
[245] K. Etzahlan, S. Siddiqui, and K. Shaalan, "Evaluating Arabic parser and recommending improvements," in Proc. Int. Conf. Adv. Intell. Syst. Inform. Cham, Switzerland: Springer, 2018, pp. 417–428.
[246] A. Alsudais, W. Alotaibi, and F. Alomary, "Similarities between Arabic dialects: Investigating geographical proximity," Inf. Process. Manage., vol. 59, no. 1, Jan. 2022, Art. no. 102770.
[247] Y. Marton and I. Zitouni, "Transliteration normalization for information extraction and machine translation," J. King Saud Univ. Comput. Inf. Sci., vol. 26, no. 4, pp. 379–387, Dec. 2014.
[248] K. Meftouh, N. Bouchemal, and K. Smaïli, "A study of a non-resourced language: An Algerian dialect," in Proc. Spoken Lang. Technol. Under-Resourced Lang., Jan. 2012, pp. 125–132.
[249] S. Sharoff, R. Rapp, and P. Zweigenbaum, "Extraction of parallel sentences," in Building and Using Comparable Corpora for Multilingual Natural Language Processing. Cham, Switzerland: Springer, 2023, pp. 39–48.
[250] S. Jafaritazehjani, G. Lecorvé, D. Lolive, and J. D. Kelleher, "Style as sentiment versus style as formality: The same or different?" in Proc. 30th Int. Conf. Artif. Neural Netw. Mach. Learn. (ICANN), Bratislava, Slovakia. Cham, Switzerland: Springer, Jan. 2021, pp. 487–499.
[251] H. Alhamed and K. Haddar, "Toward building a bilingual dictionary for Libyan dialect-modern standard Arabic machine translation," in Proc. KDIR, 2022, pp. 75–81.
[252] B. Talafha, A. Abuammar, and M. Al-Ayyoub, "Atar: Attention-based LSTM for arabizi transliteration," Int. J. Electr. Comput. Eng., vol. 11, no. 3, p. 2327, Jun. 2021.
[253] M. Wu, T.-T. Vu, L. Qu, G. Foster, and G. Haffari, "Adapting large language models for document-level machine translation," 2024, arXiv:2401.06468.

SHADI I. ABUDALFA received the B.Sc. and M.Sc. degrees in computer engineering from the Islamic University of Gaza (IUG), Palestine, in 2007 and 2010, respectively, and the Ph.D. degree in computer science and engineering (CSE) from the King Fahd University of Petroleum and Minerals (KFUPM), in 2018. He has a strong teaching experience through his work as an Assistant Professor with the University College of Applied Sciences, Palestine. He is currently a Researcher with the SDAIA-KFUPM Joint Research Center for Artificial Intelligence (JRC-AI). From July 2003 to August 2004, he was a Research Assistant with Projects and Research Laboratory, IUG. During the same period, he was a Teaching Assistant with the Faculty of Engineering, IUG. He serves as a reviewer through various international journals and conferences. His current research interests include data mining, activity recognition, Arabic natural language processing, and sentiment analysis.

FAHAD I. ABDU was born in Dogowa, Kano, Nigeria. He received the B.Eng. degree in electrical engineering from Bayero University Kano (BUK), Kano, in 2011, the M.Sc. degree in electrical and electronics engineering from Mevlana University, Konya, Türkiye, in 2014, and the Ph.D. degree in communication and information systems from Xiamen University (XMU), Xiamen, China, in 2022. He is currently a Postdoctoral Researcher with the SDAIA-KFUPM Joint Research Center for Artificial Intelligence, King Fahd University of Petroleum and Mineral Resources (KFUPM). He is also a Lecturer with the Department of Electronics and Telecommunications Engineering, Ahmadu Bello University (ABU), Zaria, Nigeria. His current research interests include radar signal processing, multisensory fusion, computer vision, artificial intelligence (AI), and Arabic natural language processing (NLP).

MAAD M. ALOWAIFEER received the B.Sc. and M.Sc. degrees in electrical engineering from King Fahd University of Petroleum and Minerals (KFUPM), Dhahran, Saudi Arabia, in 2013 and 2015, respectively, and the M.Sc. degree in operations research and the Ph.D. degree in electrical and computer engineering from Georgia Institute of Technology, Atlanta, USA, in 2020 and 2021, respectively. He is currently an Assistant Professor with the Electrical Engineering Department, KFUPM. His research interests include optimization, artificial intelligence, and their applications in the energy sector.