applied sciences

MDPI

Article

Domain Adaptation for Arabic Machine Translation: Financial Texts as a Case Study

Emad A. Alghamdi [1,2,*,†], Jezia Zakraoui [2,†] and Fares A. Abanmy [2]

[1] Center of Excellence in AI and Data Sciences, King Abdulaziz University, Jeddah 21589, Saudi Arabia
[2] ASAS AI Lab, Riyadh 13518, Saudi Arabia
* Correspondence: eaalghamdi@kau.edu.sa
† These authors contributed equally to this work.

Abstract: Neural machine translation (NMT) has shown impressive performance when trained on large-scale corpora. However, generic NMT systems have demonstrated poor performance on out-of-domain translation. To mitigate this issue, several domain adaptation methods have recently been proposed which often lead to better translation quality than genetic NMT systems. While there has been some continuous progress in NMT for English and other European languages, domain adaption in Arabic has received little attention in the literature. The current study, therefore, aims to explore the effectiveness of domain-specific adaptation for Arabic MT (AMT), in yet unexplored domain, financial news articles. To this end, we developed a parallel corpus for Arabic-English (AR-EN) translation in the financial domain to benchmark different domain adaptation methods. We then fine-tuned several pre-trained NMT and Large Language models including ChatGPT-3.5 Turbo on our dataset. The results showed that fine-tuning pre-trained NMT models on a few well-aligned in-domain AR-EN segments led to noticeable improvement. The quality of ChatGPT translation was superior to other models based on automatic and human evaluations. To the best of our knowledge, this is the first work on fine-tuning ChatGPT towards financial domain transfer learning. To contribute to research in domain translation, we made our datasets and fine-tuned models available.

Keywords: machine translation; Arabic MT; domain adaptation; financial domain

check for updates

Citation: Alghamdi, E.A.; Zakraoui, J.; Abanmy, F.A. Domain Adaptation for Arabic Machine Translation: Financial Texts as a Case Study. *Appl. Sci.* **2024**, *14*, 7088. https://doi.org/10.3390/app14167088

Academic Editor: Tobias Meisen

Received: 30 June 2024
Revised: 24 July 2024
Accepted: 1 August 2024
Published: 13 August 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In recent years, the rapid advancement of deep learning techniques and their adaptation in machine translation has made a great stride in many translation tasks. Neural Machine Translation (NMT) systems trained on a large-scale corpus, have demonstrated impressive performance in translating generic language. However, NMT models tend to perform poorly on out-of-domain data [1], especially if the target domain has a distinctive style and vocabulary [2]. Accordingly, an NMT model trained on exclusively medical texts is unlikely to achieve accurate performance on financial or news data. To address this problem, researchers have proposed different domain adaptation approaches and techniques that seem to improve the quality of NMT systems on out-of-domain data [3–5].

There are many MT models, systems and tools for translating Arabic texts in the literature; however, the quality of the translation is poor, especially for out-of-domain texts [6,7]. A key technical challenge related to AMT arises from the lack of available bilingual datasets for out-of-domain texts that can be used as standard benchmarks to conduct unified experiments. In fact, researchers tend to collect datasets according to their specific domains and try to resolve the linguistic issues for Arabic, based on custom datasets such as in the domain of news [8,9], ignoring thereby many other domains. Other technical issues such as out-of-vocabulary (OOV) and very long sentences also make MT more challenging [1]. To address these challenges, researchers have proposed different techniques, including, for example, BPE [10], character-level BPE variant [11], hybrid

techniques [12], and mixed fine-tuning [6]. However, domain robustness remains an unsolved problem and there is a need for further research in this area [13]. This is especially true for the Arabic language. Existing domain adaptation research has only focused on news [14] and medical [7] domains, no prior study, to the best of our knowledge, has been conducted on the financial domain.

To alleviate the issue of translation mismatch related to out-of-domain texts, the authors in [14] studied the performance of NMT systems under morphology-based and frequency-based tokenization schemes and BPE on in-domain data. They evaluated their best-performing models on out-of-domain data yielding significant improvements of 37.96% in BLEU score [15]. Ref. [7] proposed a method for domain-specific data augmentation for MT to tackle the issue with a small bilingual dataset. They employed mixed fine-tuning to train models that significantly improve the translation of in-domain texts. Their method achieved improvements of approximately 5–6 BLEU and 2–3 BLEU, respectively, on the Arabic-to-English and English-to-Arabic language pairs.

While a lot of research in domain adaptation in MT for other language pairs like [6], ref. [16] exists which focuses on synthetic data generation and multiple other techniques like checkpoint averaging [6], only one work [7] investigated the same for AMT, but only for a medical domain. Therefore, this research aims to fill the gap in evaluating different MT settings and investigate domain adaption for financial texts. Our contributions are the following:

- We introduce the first AR-EN parallel corpus in the financial domain.
- We compare the effectiveness of different adaption methods and data augmentation approaches for limited domain data.
- We fine-tuned several models and made them publicly available to the research community at https://huggingface.co/asas-ai/ (accessed on 20 July 2024).
- Our work is the first to fine-tune the GPT3.5 model and evaluate its capability for domain adaption.

2. Background

2.1. Neural Machine Translation

NMT models based on deep neural networks (DNN) have been proposed in early NMT research [17]. A DNN-based NMT model employs a neural network system to perform the required machine translation tasks using an encoder-decoder network [18]. The encoder neural network inputs and encodes a source language sentence into a fixed-length vector in each hidden state. Then, given the final hidden state of the encoder, the decoder does the reverse work by transforming the hidden state vector to the target sentence word by word. A translation probability of a source sentence is modeled into the target sentence. Given a source sentences $S = \{s_1, s_2, ..s_n\}$ and a target sentence $T = \{t_1, t_2, ..t_n\}$, the encoder encodes all the words from the source sentence S into a set of hidden states $(h_1, h_2, ..h_n)$ and passes the fixed-size vector v, which represents the source sentence, to the decoder. The translation probability with a single neural network is given by the following formula [19]:

$$P(S) = \prod_{i=1}^{n} P(t_{<i}, S) \quad (1)$$

where $t < i$ stands for the sequence preceding the i-th target word. Hence, each predicted word t_i is based on the previously predicted word t_{i-1} and the previous hidden states h_{i-1}. However, when the sentences become long the performance deteriorates. This limitation is due to the limited feature representation ability in a fixed-length vector [17]. To overcome this issue and to provide additional word alignment information in translating long sentences, Bahdanau et al. [20] introduced the idea of the attention mechanism. Concretely, the attention mechanism is an intermediate component between the encoder and decoder, which can help to determine the word alignment dynamically. The decoder pays attention to input or to any part of the input sentence. Attention is calculated using each encoder output and the current hidden state, resulting in a vector of the same size as

the input sequences using score functions [20]. There are three different architectures for constructing NMT, namely Recurrent neural network (RNN), Convolution neural network (CNN), and Self-attention-based Transformer.

The use of RNN-based models has demonstrated good-quality translation results. This type of network is composed of an encoder and decoder with similar uses of sequence-to-sequence learning. Multiple variants of RNN architectures include, i.e., LSTM [21], BiLSTM [20] and GRU [22].

The second approach to developing NMT systems is based on convolution neural network (CNN) architecture. Work using CNN has generally reported good results, especially for word-based MT [23]. This work applied a convolution layer on the bottom of the recurrent layer which hinders the performance. The bottleneck was handled by implementing the fully convolutional model as suggested by [24]. The performance and accuracy were improved with a number of models; word-based [25], character-based [11], and recently with attention [26].

Recently, the use of transformers has resulted in well-performing machine translation systems. This model is a sequence-to-sequence model [27], which consists of a stack of layers. Each layer first utilizes self-attention to extract information from the whole sentence and then follows a point-wise feed-forward network to provide non-linearity. The novel idea of self-attention is to extend the mechanism to the processing of input sequences and output sentences as well. In general form, the Transformer attention function uses three vectors: queries (Q), keys (K) and values (V). The output is a weighted sum of values, where weights are computed by a similarity score between n query vectors and m keys [27]. The attention is defined as follows:

$$Attention(Q, K, V) = softmax(score(Q, K))V \quad (2)$$

where score Q, K is an n × m matrix of similarity scores. A straightforward choice for score Q, K proposed by Luong et al. [28] is the dot product, i.e., score (Q, K) = QK. The softmax function normalizes over the columns of that matrix so that the weights for each query vector sum up to one. There are many variants in the implementation of attention-based models which are classified into two broad categories, global and local attention discussed in detail in this survey [17].

The current state-of-the-art NMT models [29] rely on the Transformer model [27] and multiple attention mechanism [20]. However, transformer-based language models such as Bidirectional Encoder Representation from Transformers (BERT) [30] expand the function of attention to encompass the main task. It uses self-attention, which is applied to two states within the same sequence, as the foundation for sequence representations rather than an RNN. For the Arabic language, two transformer-based language models have been developed so far; notably AraBERT [31] and GigaBERT [32]. Both models aim at solving a masked language-modeling task in order to correctly predict a masked word from its context. Besides, these models aim at resolving a next sentence prediction task especially to decide whether two sentences are consecutive or not.

2.2. Domain-Specific MT

Domain translation is a challenging task due to the fact that language varies across different domains, genres, and styles. For example, texts in the financial domain often contain specific terminologies and jargon that may not be extensively used in legal or health domains. Therefore, researchers have proposed different methods to improve the quality of translations in domains such as medical and biomedical [7,33,34], legal [35], and financial texts [36]. Several domain adaptation approaches have been proposed (for more comprehensive survey see [3]). Domain adaptation methods can intervene in various stages of NMT system design, training and use and can be classified into three main categories: data-centric methods, architecture-centric adaptation methods, and inference schemes for adaptation [3]. In data-centric methods, the objective is to select or generate appropriate in-domain data. Large generic monolingual data can be filtered to select

domain-representative datasets based on some unique characteristics of the target domain. However, selecting a small in-domain dataset may be more domain-relevant, but the impact of any deviation from the target domain will be magnified [37]. Another approach is to construct partially synthetic bilingual training corpora by forward- or back-translation. Ref. [38] observed that models trained exclusively on back translations can perform similarly to models trained on natural data. Recently, the use of pre-trained large language models (LLMs) to generate large amounts of synthetic data at very low cost has emerged to be an effective approach [7].

Architecture-centric adaptation typically involves adding trainable parameters to pre-trained models to avoid training models from scratch. A common approach is to fine-tune an existing well-performing NMT model on small in-domain data. Extensive fine-tuning can lead to catastrophic forgetting. Ref. [39] proposed mixed-fine tuning which involves two steps: (1) training an NMT model on out-of-domain data until convergence and then (2) fine-tuning the NMT model from step 1 on a mix of in-domain and out-of-domain data (by oversampling the in-domain data) until convergence. Mixed-fine tuning approaches can be helpful to prevent two major issues notably overlooking the specificity of each domain [1] and forgetting previously learned knowledge when exposed to the new training examples as reported in [40].

Lastly, the inference schemes for adaptation develop a separate NMT model for each domain and combine them at inference time.

2.3. Domain-Adaptation in Arabic MT

The development of Arabic MT systems has gone through different stages, including rule-based systems [41,42], statistical MT [43], and more recently neural MT systems [44]. Ref. [45] conducted a comprehensive survey of Arabic MT systems and the unique challenges in Arabic MT.

Arabic is one of the official six languages adopted by the United Nations and it is spoken by 400 million people in the Middle East, North Africa, and many other parts of the world. Arabic is a Semitic language and it is notoriously difficult for MT due to its linguistic characteristics [45,46]. First, Arabic has a rich and complex morphology which is substantially different from English or other western languages [47]. Second, Arabic has long and short vowels. While the long vowels are represented by letters, the short vowels are marked by diacritic signs placed above or below the letters. However, the use of diacritic signs is not compulsatory in Arabic and hence they are rarely used in informal writing. Therefore, it is hard to identify the correct sense of a word, especially when sufficient context is not provided. Third, variation among different Arabic dialects has always been problematic for AMT. Furthermore, the Arabic language used in social media varies considerably from Modern Standard Arabic (MSA). These aspects of the Arabic language pose series challenges for Arabic MT.

In addition to the aforementioned issues, there is a lack of high-quality parallel corpora of sufficient size for training or fine-tuning Arabic MT for different domains. It is commonly known that NMT systems do not perform well in domain-specific translation, especially in low-resource languages [1]. Addressing these challenges, some researchers have turned to domain adaption methods to develop domain-specific Arabic MT systems. For example, ref. [7] proposed the use of pre-trained LMs and back-translation for domain-specific data augmentation for MT.

Furthermore, current Arabic MT research has primarily focused on the translation of limited domains such as news and official texts, whilst few attempts focus on domain-specific translation such as medical domain [48]. Specifically, most of the parallel data available to the researcher was limited to texts produced by international organizations, and parliamentary debates [33]. Unfortunately, existing single-domain AMT methods do not work well for multiple domains. Thus, multi-domain NMT approaches are more in demand to tackle this limitation.

To recap, previous research has shown that domain adaptation leads to better translation quality than general NMT. While there has been considerable progress in general MT from Arabic to English [5,49], less work has been conducted on adapting models to specific domains such as medical domains [7], but to the best of our knowledge no work investigated the adaptability in financial texts. Since there is relatively little work on Arabic domain adaptation, the primary objective of this research is to explore the different effectiveness of domain translation methods, a yet unexplored domain, financial domain. To this end, this work aims to fine-tune several Transformer NMT models and LLM and perform cross-domain testing and evaluation to gain some insights into model robustness against domain changes.

3. Methodology

This section gives an overview of the methods and algorithms for AMT domain adaptation using LLM models. First, information about the collected bilingual dataset used is given which we refer to as the authentic dataset, then our approach is presented, and lastly, the metrics we used for evaluation are described.

3.1. Approach

In this work, we investigate mainly two methods to augment our in-domain data for the domain of financial news and propose approaches to leverage pre-trained LLMs for domain-specific data generation for this MT task. Concerning domain-specific data generation, we start with synthetic data generation to augment our authentic sentences for Arabic. Then, to obtain the parallel data in English, we apply forward translation from the Arabic synthetic sentences into English.

3.1.1. Synthetic Data Generation

Synthetic data generation for data augmentation has been used in domain translation due to the scarcity of domain-specific datasets that are suitable for training large models. Ref. [7] proposed the use of state-of-the-art large language models to general unlimited new sentences in the source language and then back-translating in the target language. Recent studies explored the use of ChatGPT for generating new parallel sentences. However, in this study [50], the authors showed that the performance of ChatGPT for Arabic shows inferior performance compared to the finetuned AraT5. In our case, we leverage a pipeline of different models. We start with AraGPT2 [51] and gpt2 [52] as text generation models for Arabic and English to create synthetic pairs for (AR-EN) and (EN-AR), respectively. For Arabic, we use titles only from the collected authentic dataset as text prompts to generate corresponding long-form text using AraGTP2 hosted on HuggingFace (https://huggingface.co/aubmindlab/aragpt2-base, accessed on 13 March 2024). Then, we use mT5 [53] a summarization model hosted on HuggingFace (https://huggingface.co/eslamxm/mt5-base-arabic, accessed on 20 March 2024) to summarize the generated bunches of texts to obtain short summaries that will serve as generated titles. After that, we apply forward translation from Arabic to English using the OPUS-MT model published on HuggingFace (https://huggingface.co/Helsinki-NLP/opus-mt-ar-en, accessed on 20 March 2024). Figure 1 shows the case of augmenting the authentic dataset with AR-EN synthetic pairs using this method. The same pipeline applies to English to obtain EN-AR synthetic pairs.

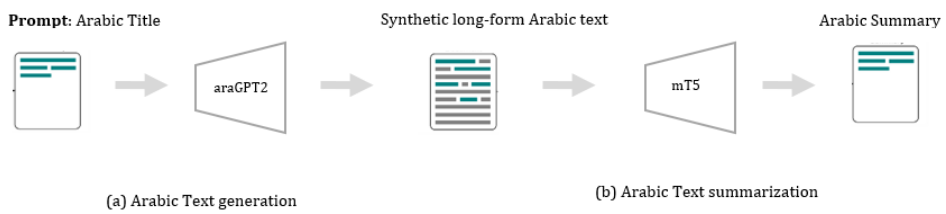

Figure 1. Data augmentation pipeline using (**a**) Arabic text generation and (**b**) Arabic text summarization.

3.1.2. Back-Translation

A common approach to augment domain data is the use of back-translation when there is abundant data in the target domain [1,38]. We use a pre-trained machine translation model [54] published on HuggingFace (https://huggingface.co/Helsinki-NLP/opus-mt-en-ar, accessed on 20 July 2024) for back-translation. The back-translation is applied on both sides of generated summaries and titles, namely on the long-form text (which serves as an article) as well as on the summarized form (which serves as a title) into the respective target language. In the end, we pair the generated summaries as well as the long-form text to serve as the title and article, respectively. The same pipeline applies to English as the target language. Figure 2 shows the case of augmenting the authentic dataset with EN-AR back-translated pairs.

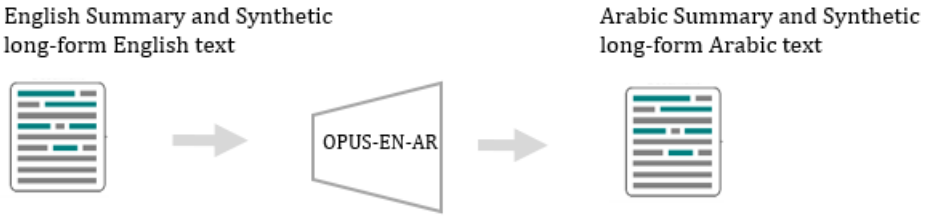

Figure 2. Data augmentation pipeline using back-translation from EN to AR.

3.2. Experiment Setup

3.2.1. Datasets

For fine-tuning in domain-specific MT models, we collected a dataset from different online resources for the pair AR-EN. As shown in Table 1 most of the data are collected from the Capital Markets Authority (CMA) yielding a total of 7560 AR-EN pairs. Note, that we consider titles (3780 AR-EN pairs) and articles (3780 AR-EN pairs). Additionally, we augmented our dataset with synthetic data as well as back-translated data. This step augmented the authentic dataset by 12,318 and 12,000 AR-EN sentence pairs as synthetic and back-translated data, respectively.

Table 2 shows the breakdown of the segments in our dataset. We randomly sampled 1000 segments from the authentic dataset to serve as test data for all models. Additionally, we randomly sampled 1000 segments for building the development for both models notably for OPUS (bt-big) and NLLB. However, for fine-tuning ChatGPT, we randomly sampled 2000 pairs each for each setup.

Table 1. Authentic dataset statistics.

Source	Articles	Titles	Sentences
Tadawul	569	569	2544
Capital Markets Authority	2320	2320	8351
Eye of Riyadh	891	891	1877
Total	3780	3780	15,771

Table 2. Authentic dataset split and augmented data count.

Language Pair	Type	Fine-Tuning	Dev	Test
AR-EN	Authentic	5560	1000	1000
AR-EN	Synthetic	11,318	1000	-
EN-AR	Back-translated	11,000	1000	-

3.2.2. Nmt Pre-Trained Models

Our generic NMT pre-trained models use different Transformer architectures; however, we have implemented the fine-tuning objective using the huggingface NMT transformer (a sequence-to-sequence version in the Transformers library) procedure. We adopt a common Seq2Seq architecture mainly composed of Encoder and Decoder network. The fine-tuning procedure is summarized as follows:

1. Load a pre-trained model with the corresponding Tokenizer.
2. Tokenize the training, validation and test data into subwords (sentencePiece) so that training, validation and testing data will be truncated and tokenized.
3. Instantiate the loaded model using Seq2seq Trainer (https://huggingface.co/docs/transformers/en/main_classes/trainer#transformers.Seq2SeqTrainer, accessed on 1 June 2024) from huggingface while setting the AdamW (Adam with Weight Decay) optimizer with appropriate hyperparameters.
4. Fine-tune the loaded model using backpropagation while minimizing cross-entropy.
5. Save the fine-tuned model on the huggingFace hub.
6. Load the fine-tuned model for inference, feed the tokenized test data and decode the outputs to obtain the translations.

For Fine-tuning and inference, we use beam size 4 and batch size 16, on a GPU T4-15 GB (Google Colab). Further, we use ChatGPT as a baseline with zero-shot learning.

OPUS (bt-big): We use OPUS [55] models from the Tatoeba-Challenge, specifically the models augmented with back-translated data of Wikimedia content and trained with Transformer-Big architecture. Here we picked the Helsinki-NLP/opus-mt-ar-en checkpoint. For tokenization, we instantiate our tokenizer which is based on SentencePiece [56] with the AutoTokenizer.from_pretrained method. This ensures that the tokenizer corresponds to the model architecture we want to use.

NLLB: No-Language-Left-Behind (NLLB) [57] is a multilingual model which supports 200 languages with a massive size Transformer. Fine-tuning is carried out on NLLB using its distilled version facebook/nllb-200-distilled-600M checkpoint. For tokenization, we instantiate a multilingual model provided by NLLB for tokenization with the NllbTokenizerFast.from_pretrained method. This ensures that the tokenizer corresponds to the model architecture we are using.

ChatGPT3.5: We use the ChatGPT-3.5-turbo model via its official API (https://chat.openai.com, accessed on 20 August 2023) which powers ChatGPT. Here, we prepare our dataset in the format that is accepted by the API functions. In particular, we convert the AR-EN pairs into the Prompt template for sentence-level translation as recommended in the OpenAI playground for sentence-level translation tasks. In order to avoid errors, we truncate all the sentence pairs to a max size of 4290 characters before sending the request. Moreover, we set the size of the total tokens to about 378,460 tokens due to limit rate costs. For this model, we formatted the requests with the system message first 'You are a professional translator in the financial domain. Translate the following Arabic sentence: ar_en into English' followed by user content messages, where ar_en represents the AR-EN pairs.

3.2.3. Settings

Before starting the experiments, we considered the following three setups for fine-tuning the models on the domain-specific dataset. Next, Section 4 will discuss the results and findings.

Setup 1 (baseline models): We consider pre-trained NMT models evaluated on our cleaned authentic test split containing 1000 AR-EN sentence pairs. Our baseline NMT models use the OPUS (bt-big) (https://github.com/Helsinki-NLP/Tatoeba-Challenge/tree/master/models/ara-eng, accessed on 12 March 2024) [54], NLLB 600 M (https://huggingface.co/facebook/nllb-200-distilled-600M, accessed on 12 March 2024) [57] and

ChatGPT-3.5 (https://platform.openai.com/docs/guides/gpt/chat-completions-api, accessed on 12 March 2024).

Setup 2 (fine-tuning with authentic data): For fine-tuning, we have initialized the transformer models with the trained weights of the baselines. We use our authentic dataset with the splits shown in Table 2. We have kept all the hyperparameters identical. The models have been fine-tuned until convergence over the validation set. At test time, the respective test set from the authentic dataset is used for this setup as well. Again, all the metrics are reported.

Setup 3 (fine-tuning with augmented data): Similar to the previous setup, we have initialized the transformer models with the trained weights of the baselines. However, here we use our authentic dataset augmented with the respective data with the splits shown in Table 2. Basically, we augment the authentic dataset with back-translated data and shuffle it. The same applies to synthetic data. This step yields two versions of fine-tuning, one using the former and one using the latter. The models have been fine-tuned until convergence over the validation set. At test time, the test set from the authentic dataset is used for this setup as well while also reporting all metrics.

3.3. Metrics

As performance measures, we report the spBLEU score [15] which uses a SentencePiece tokenizer, chrF [58], TER [59], and are implemented in sacrebleu (https://github.com/mjpost/sacreBLEU, accessed on 2 June 2023). Additionally, we compute COMET [60] that was proposed recently by taking advantage of cross-lingual pre-trained LMs using knowledge from both source and target languages. COMET makes a prediction score that correlates with human judgment [60]. For our experiments, we adopt the official COMET implementation (https://github.com/Unbabel/COMET, accessed on 2 June 2023). For COMET, we use the reference-based Estimation model wmt20-comet-da, trained based on Direct Assessment (DA) and used Quality Estimation (QE). Another score that correlates with human evaluation BERTScore [61] is also computed. Including different metrics in the evaluation allows us to test the models on metrics different from those used for training.

4. Results and Discussion

This section elaborates on our automatic and human evaluations and discusses the results. We also provide a preliminary comparison of the models' performance on domain-specific MT as baseline models and as fine-tuned models. Therefore, we report if they can perform robustly well on domain-specific or even noisy sentences from our collected dataset. Specifically, we focus on the translation robustness of the models in the translation of Arabic financial news. Table 3 shows the main results of the respective testset. The ↑ and ↓ symbols in the tables indicate which values are better. We analyze the translation outputs by comparing the MT evaluation metrics in each setup. Visual plots of the models' performances in each of the setup are presented in Appendix A.

Table 3. MT evaluation scores and human evaluation for AR-EN Test dataset (1000 pairs). The best scores are in **bold**.

	Model	spBLEU ↑	chrF ↑	TER ↓	COMET ↑	BERTScore ↑	Human ↑
1	OPUS (bt-big) [1]	14.58	43.93	79.59	3.89	0.89	-
	NLLB 600 M [2]	14.38	42.17	77.58	2.98	0.89	-
	ChatGPT-3.5	**26.13**	**60.98**	**66.83**	**33.7**	**0.91**	-
2	OPUS (bt-big) FT [3]	48.83	65.11	53.18	51.12	**0.95**	2.7
	NLLB 600M FT [4]	43.43	61.01	54.65	**52.10**	0.94	2.81
	ChatGPT-3.5 FT	**51.15**	**71.28**	**46.47**	42.90	0.94	3.1

Table 3. *Cont.*

	Model	spBLEU ↑	chrF ↑	TER ↓	COMET ↑	BERTScore ↑	Human ↑
3	OPUS (bt-big) FT-BMT [5]	**47.56**	64.53	**54.30**	**57.21**	**0.95**	2.94
	OPUS (bt-big) FT-S [6]	40.67	57.87	60.46	49.71	0.94	2.67
	NLLB 600M FT-BMT [7]	43.38	60.92	54.63	52.77	0.94	2.67
	NLLB 600M FT-S [8]	40.77	58.26	57.48	49.44	0.94	2.85
	ChatGPT-3.5 FT-BMT	45.07	**67.64**	55.07	33.55	0.93	2.93
	ChatGPT-3.5 FT-S	34.67	62.93	70.29	23.03	0.91	2.77

BMT = Back MT, FT = Fine-tuned, S = synthetic, the direction of the arrows indicate better performance [1] https://huggingface.co/Helsinki-NLP/opus-mt-ar-en (accessed on 2 June 2024); [2] https://huggingface.co/facebook/nllb-200-distilled-600M (accessed on 2 June 2024); [3] https://huggingface.co/asas-ai/opus-mt-ar-en-finetuned-ar-to-en (accessed on 2 June 2024); [4] https://huggingface.co/asas-ai/nllb-200-distilled-600M-finetuned-ar-to-en (accessed on 2 June 2024); [5] https://huggingface.co/asas-ai/opus-mt-ar-en-finetuned_augmented_MT-ar-to-en (accessed on 2 June 2024); [6] https://huggingface.co/asas-ai/opus-mt-ar-en-finetuned_augmented_synthetic-ar-to-en (accessed on 2 June 2024); [7] https://huggingface.co/asas-ai/nllb-200-distilled-600M-finetuned_augmented_MT_ar-to-en (accessed on 2 June 2024); [8] https://huggingface.co/asas-ai/nllb-200-distilled-600M-finetuned_augmented_synthetic_ar-to-en (accessed on 2 June 2024).

4.1. Automatic Evaluation

In Setup 1, OPUS and NLLB perform equally with inferior performances of around 14 and 42 for BLEU and chrF points, respectively. The TER score which is expressed as the ratio of the number of edits to the average number of words in the reference is high for the two models. Thus, it indicates that the translation is of poor quality. In terms of COMET score, both models have very poor results which means reference-based COMET may lose information from source, translation output, or reference embeddings, except for ChatGPT-3.5. But, BERTScores for all three models are high which means that they do not correlate with COMET score. In comparison, BERTScore and COMET have a significant difference in their scores. In contrast, ChatGPT-3.5 performs competitively better (BLEU 26.13) than OPUS and NLLB models. Indeed, we are not surprised by this fact which is in line with related research works [50,62,63]. However, these findings are not consistent with a previous finding [64] where the authors evaluated ChatGPT and GPT on 4000 Arabic-English pairs and found out that SoTA models like araT5 [44] outperforms ChatGPT by 19 BLEU Points. Similarly, ref. [65] found the English-to-Arabic translation of ChatGPT-3.5 was below average compared to 14 established MT systems.

When, we analyze Setup 2, as expected, fine-tuning all models on authentic data has generally helped improve the BLEU scores and other metrics as well. This finding is also in-line with other previous research [7,56]. However, ref. [62] noticed that for domain-specific translation (e.g., in the biomedical field), ChatGPT's performance degrades considerably. We attribute this behavior to the observation that ChatGPT is capable of translating our sentences better than terminologies in sentences from the biomedical domain, a very specific domain. Furthermore, we clearly, see that BLEU scores increase from 14.58 to 48.83, from 14.38 to 43.43 and from 26.13 to 51.15 for OPUS, NLLB and ChatGPT-3.5, respectively. In terms of COMET and BERTScore, both metrics had a high correlation which indicates acceptable translation outputs.

Concerning ChatGPT, even though it only used 2000 pairs of AR-EN sentences for fine-tuning, it outperforms all other models which means the MT quality of ChatGPT can easily be improved with little additional data from the language pair, a fact that has not been previously confirmed for related approaches, since this is the first work that assesses the performance of ChatGPT fine-tuned models for AR-EN MT task. Nevertheless, for English, this work [66] has shown that ChatGPT has great robust translation capabilities over related SoTA MT models. Our experimental result confirms the latter finding and shows that with a carefully prepared certain amount of fine-tuning data, this model is capable of creating acceptable translations. As for the translation robustness, results from Setup 2 suggest that ChatGPT-3.5 performs competitively well on financial news. Regarding the

human evaluation, all models in this setup reached possible and acceptable translations. We conclude that our experiment shows that providing in-domain examples to ChatGPT achieves comparable results to a SoTA model in terms of automatic and human evaluation.

In Setup 3 we fine-tune the baseline models with the augmented data in two versions, one using back-translated data and the other using synthetic data. We observe that both lexical metrics (BLEU and chrF) show consistent degradation with all models. The same applies to the TER score. For instance, for ChatGPT, the BLEU score decreased dramatically from 51.15 to 34.67 when fine-tuned on synthetic data while maintaining an acceptable score (BLEU 45.38) when fine-tuned on back-translated data. We observe that the COMET score degraded massively for ChatGPT more than for OPUS and NLLB. One explanation could be that the synthetic data may have a lot of generated tokens that are grammatically correct, but they have nonsense meaning, as we know from the current state of the generative text. This could indicate that the translation results are not close in the embedding space with the source and reference. In contrast, BERTScore maintained a good score over the two versions for all models. In this setup, OPUS (bt-big) FT (back MT) has made it the best model that provides reasonably good scores translations; however, it still lags behind the OPUS model fine-tuned on authentic data by at least 1.3 BLEU points.

Generally, the drop in performance for all models in this setup is not consistent with others' research. For instance, the authors in [7] used synthetic data in the healthcare domain and achieved improvements on the in-domain test set. In comparison, with this work, the authors applied synthetic data generation using mGPT (https://huggingface.co/sberbank-ai/mGPT, accessed on 12 March 2024) a multilingual language model. We argue that this model might have better perplexity in generated tokens compared to araGPT2. To the best of our knowledge, we did not find any research work investigating the performance of both models in regard to Arabic. We will further investigate this issue in future work. However, there are many general reasons explaining OPUS, NLLB and ChatGPT behavior in domain-specific MT, especially in the case of augmenting the dataset with synthetic data. One explanation is that the use of synthetic data may cause incorrect token choices, grammatical errors, or unnatural sentence structures to propagate into the translation outputs which make suboptimal translation outputs.

Indeed, the results of this study demonstrate the models' robust translation capabilities for in-domain adaptation. They perform well when fine-tuned on authentic data. However, we observe a discrepancy between COMET and BERTScore. For instance, ChatGPT-3.5 performs worse on augmented data yielding a lower COMET score (23.03) but still having high BERTScore (0.91). This behavior seems uncommon. One possible explanation is that COMET with reference-based translation is failing to find closeness in all three resource embeddings, whereas BERTScore is able the find closeness in the similarity between an MT output and a reference translation. This behavior encourages us to drive human evaluation a much-needed score for trustworthiness.

4.2. Human Evaluation

In addition to the automatic evaluations reported above, we decided to assess the quality of our models' translations using human evaluation. Machine translation metrics such as BLEU only measure the linguistic proximity of outputs to the gold standard of reference. On the other hand, COMET and BERTScore which overcome this aforementioned issue, still exhibit discrepancy issues. To gain further insights into these results, we conducted a human evaluation. To this end, we recruited three native speakers and domain experts (post-graduate students in finance) to rate the acceptability of 50 randomly selected sentences from the test set. Similar to [7], we conducted a bilingual evaluation, whereby the evaluators rated both the original source sentences and translations generated by the MT models. The human evaluators were asked to rate each of the sentences based on the scale proposed by [67], ranging from 1 to 4, and outlined as follows:

- 4 = Ideal: Not necessarily a perfect translation, but grammatically correct, with all information accurately transferred.

- 3 = Acceptable: Not perfect (stylistically or grammatically odd), but definitely comprehensible, and with the accurate transfer of all important information.
- 2 = Possibly Acceptable: Possibly comprehensible (given enough context and/or time to work it out); some information transferred accurately.
- 1 = Unacceptable: Absolutely not comprehensible and/or little or no information is accurately transferred.

We first asked the three human evaluators to rate one model's output and then we conducted an inter-rater reliability analysis on their ratings. The result of weighted Cohen's Kappa is 0.87. Then, we asked each evaluator to rate the outputs of the remaining models and provide justification for their responses were "Ideal" or "Unacceptable". The mean value of the raters' scores was averaged for each system, as shown in Table 3.

Overall, the outcome of human evaluation corroborates the automatic evaluation results. In Setups 1 and 2, where ChatGPT-3.5 achieved the best performance in lexical and semantic metrics, human evaluation confirmed this result with a top score of 3.1.

However, in Setup 3, even though the automatic metrics are degraded for all models, except BERTScore, the human evaluation shows that the translation quality of all models is comparable.

Thus, we find that BERTScore correlates with human judgment more than COMET which has been recently reported to correlate highly with with human judgment [68]. This finding opens a great investigation for the future into whether semantic metrics correlate with human judgment and to what extent, in particular, when ChatGPT-3.5 is applied.

5. Conclusions

In this paper, we conducted several experiments to assess the performance of pre-trained NMT and LLM like GPT-3.5 using data augmentation in the domain of Arabic financial news articles. Generally, the results obtained from these experiments are very promising. While ChatGPT shows good results using few pairs, other models need more examples and still have lower performance. We explored the effectiveness of all models using data augmentation in the financial domain and found that MT quality decreased for all the models adequately. Here, ChatGPT shows inferior performance, while OPUS still performs better on back-translated data than on synthetic data.

There are many future works that can be carried out based on the findings from this study. Firstly, we would like to explore new techniques and methods to enhance translation outputs rather than the approach of data augmentation. Secondly, we think it is valuable to integrate more high-performance automatic metrics into the comparison that take semantics into consideration in a better way than in COMET and BERTScore. Finally, we will explore novel approaches to integrate additional models or even incorporate domain-specific models for improved translation performance.

Author Contributions: Conceptualization, E.A.A. and J.Z.; methodology, E.A.A. and J.Z.; validation, E.A.A., J.Z. and F.A.A.; resources, E.A.A.; data curation, F.A.A.; writing—original draft preparation, E.A.A. and J.Z.; writing—review and editing, E.A.A., J.Z. and F.A.A.; visualization, J.Z. and F.A.A.; supervision, E.A.A.; project administration, E.A.A.; funding acquisition, E.A.A. All authors have read and agreed to the published version of the manuscript.

Funding: The research team obtained funding from the Translation Studies and Research Grants Program of the General Authority for Literature, Publishing and Translation at the Ministry of Culture in the Kingdom of Saudi Arabia to complete this research study in the field of translation for the year 2022.

Institutional Review Board Statement: The human study was approved by the Ethics Committee at the English Language Institute in King Abdulaziz University (protocol code: EA23TR2-v1, date of approval: 5 February 2023).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: We made our datasets and fine-tuned models available at https://huggingface.co/asas-ai/ (accessed on 19 July 2024).

Acknowledgments: We would like to acknowledge the support from the Center of Excellence in AI and Data Sceince and the Center of Excellence of High-Performance Computing at King Abdulaziz University.

Conflicts of Interest: Authors Emad A. Alghamdi, Jezia Zakraoui and Fares A. Abanmy was employed by the company ASAS AI Lab. The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Appendix A

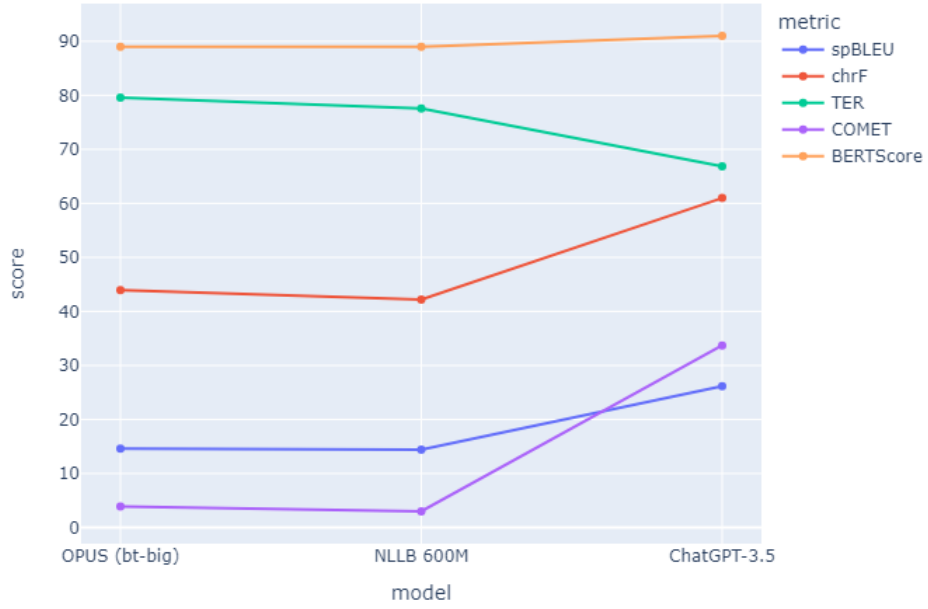

Figure A1. Plotting the models' performance in setup 1.

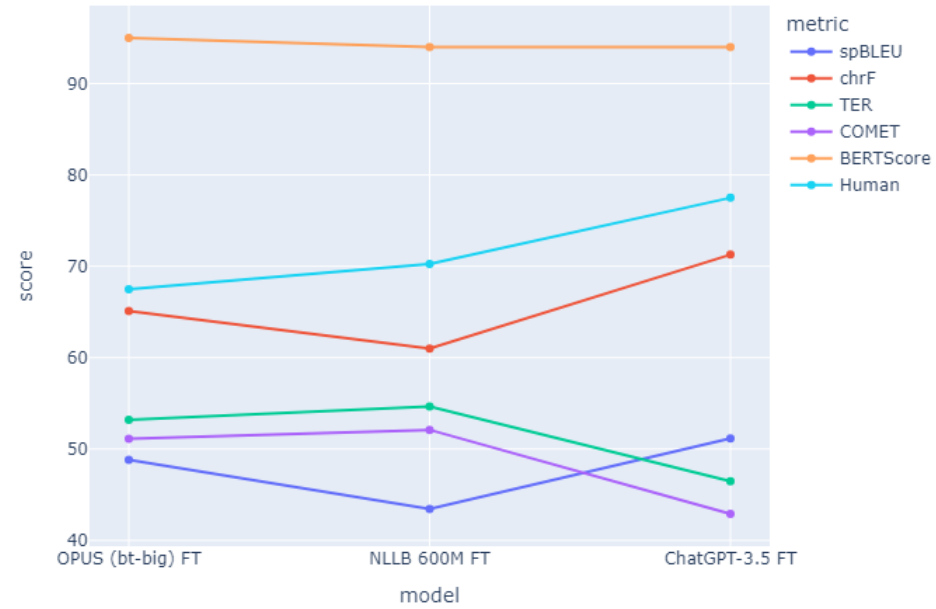

Figure A2. Plotting the models' performance in setup 2.

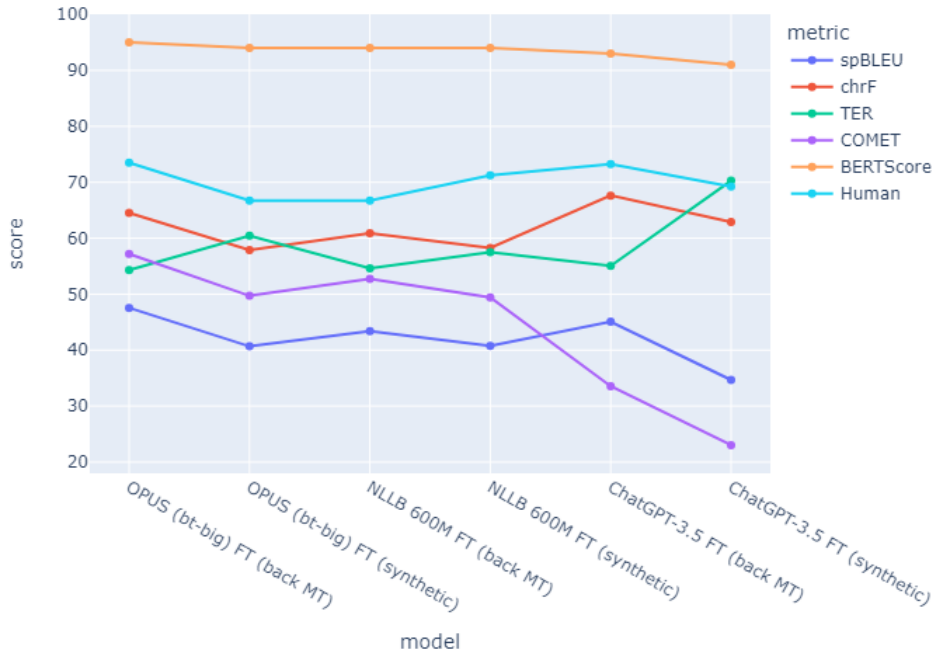

Figure A3. Plotting the models' performance in setup 3.

References

1. Koehn, P.; Knowles, R. Six challenges for neural machine translation. *arXiv* **2017**, arXiv:1706.03872.
2. Daumé Iii, H.; Jagarlamudi, J. Domain adaptation for machine translation by mining unseen words. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, OR, USA, 19–24 June 2011; pp. 407–412.
3. Saunders, D. Domain adaptation and multi-domain adaptation for neural machine translation: A survey. *J. Artif. Intell. Res.* **2022**, *75*, 351–424. [CrossRef]
4. Chu, C.; Wang, R. A survey of domain adaptation for machine translation. *J. Inf. Process.* **2020**, *28*, 413–426. [CrossRef]
5. Moslem, Y.; Haque, R.; Way, A. Adaptive machine translation with large language models. *arXiv* **2023**, arXiv:2301.13294.
6. Popel, M.; Tomkova, M.; Tomek, J.; Kaiser, Ł.; Uszkoreit, J.; Bojar, O.; Žabokrtskỳ, Z. Transforming machine translation: A deep learning system reaches news translation quality comparable to human professionals. *Nat. Commun.* **2020**, *11*, 4381. [CrossRef] [PubMed]
7. Moslem, Y.; Haque, R.; Kelleher, J.D.; Way, A. Domain-Specific Text Generation for Machine Translation. *arXiv* **2022**, arXiv:2208.05909.
8. Hatem, A.; Omar, N. Syntactic reordering for Arabic-English phrase-based machine translation. In *Database Theory and Application, Bio-Science and Bio-Technology*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 198–206.
9. Almahasees, Z.M. Assessment of Google and Microsoft Bing translation of journalistic texts. *Int. J. Lang. Lit. Linguist.* **2018**, *4*, 231–235. [CrossRef]
10. Sennrich, R.; Haddow, B.; Birch, A. Neural machine translation of rare words with subword units. *arXiv* **2015**, arXiv:1508.07909.
11. Costa-Jussa, M.R.; Fonollosa, J.A. Character-based neural machine translation. *arXiv* **2016**, arXiv:1603.00810.
12. Luong, M.T.; Manning, C.D. Achieving open vocabulary neural machine translation with hybrid word-character models. *arXiv* **2016**, arXiv:1604.00788.
13. Müller, M.; Rios, A.; Sennrich, R. Domain robustness in neural machine translation. *arXiv* **2019**, arXiv:1911.03109.
14. Oudah, M.; Almahairi, A.; Habash, N. The impact of preprocessing on Arabic-English statistical and neural machine translation. *arXiv* **2019**, arXiv:1906.11751.
15. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. BLEU: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, Philadelphia, PA, USA, 6–12 July 2002; pp. 311–318.
16. Bapna, A.; Arivazhagan, N.; Firat, O. Simple, scalable adaptation for neural machine translation. *arXiv* **2019**, arXiv:1909.08478.
17. Yang, S.; Wang, Y.; Chu, X. A survey of deep learning techniques for neural machine translation. *arXiv* **2020**, arXiv:2002.07526.
18. Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to sequence learning with neural networks. In Proceedings of the 27th International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014.
19. Johnson, M.; Schuster, M.; Le, Q.V.; Krikun, M.; Wu, Y.; Chen, Z.; Thorat, N.; Viégas, F.; Wattenberg, M.; Corrado, G.; et al. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Trans. Assoc. Comput. Linguist.* **2017**, *5*, 339–351. [CrossRef]
20. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv* **2014**, arXiv:1409.0473.

21. Zhou, J.; Cao, Y.; Wang, X.; Li, P.; Xu, W. Deep recurrent models with fast-forward connections for neural machine translation. *Trans. Assoc. Comput. Linguist.* **2016**, *4*, 371–383. [CrossRef]
22. Ataman, D.; Aziz, W.; Birch, A. A latent morphology model for open-vocabulary neural machine translation. *arXiv* **2019**, arXiv:1910.13890.
23. Meng, F.; Lu, Z.; Wang, M.; Li, H.; Jiang, W.; Liu, Q. Encoding source language with convolutional neural network for machine translation. *arXiv* **2015**, arXiv:1503.01838.
24. Gehring, J.; Auli, M.; Grangier, D.; Dauphin, Y.N. A convolutional encoder model for neural machine translation. *arXiv* **2016**, arXiv:1611.02344.
25. Kalchbrenner, N.; Espeholt, L.; Simonyan, K.; Oord, A.v.d.; Graves, A.; Kavukcuoglu, K. Neural machine translation in linear time. *arXiv* **2016**, arXiv:1610.10099.
26. Gehring, J.; Auli, M.; Grangier, D.; Yarats, D.; Dauphin, Y.N. Convolutional sequence to sequence learning. In Proceedings of the International Conference on Machine Learning, PMLR, Sydney, Australia, 6–11 August 2017; pp. 1243–1252.
27. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017.
28. Luong, M.T.; Pham, H.; Manning, C.D. Effective approaches to attention-based neural machine translation. *arXiv* **2015**, arXiv:1508.04025.
29. Stahlberg, F. Neural machine translation: A review. *J. Artif. Intell. Res.* **2020**, *69*, 343–418. [CrossRef]
30. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
31. Antoun, W.; Baly, F.; Hajj, H. Arabert: Transformer-based model for Arabic language understanding. *arXiv* **2020**, arXiv:2003.00104.
32. Lan, W.; Chen, Y.; Xu, W.; Ritter, A. Gigabert: Zero-shot transfer learning from english to arabic. In Proceedings of the 2020 Conference on Empirical Methods on Natural Language Processing (EMNLP), Online, 16–20 November 2020.
33. Abdul-Rauf, S.; Kiani, K.; Zafar, A.; Nawaz, R. Exploring transfer learning and domain data selection for the biomedical translation. In Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2), Florence, Italy, 1–2 August 2019; pp. 156–163.
34. Liu, B.; Huang, L. ParaMed: A parallel corpus for English–Chinese translation in the biomedical domain. *BMC Med. Inform. Decis. Mak.* **2021**, *21*, 258. [CrossRef]
35. Martínez-Domínguez, R.; Rikters, M.; Vasiļevskis, A.; Pinnis, M.; Reichenberg, P. Customized Neural Machine Translation Systems for the Swiss Legal Domain. In Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 2: User Track), Online, 6–9 October 2020; pp. 217–223.
36. Läubli, S.; Amrhein, C.; Düggelin, P.; Gonzalez, B.; Zwahlen, A.; Volk, M. Post-editing productivity with neural machine translation: An empirical assessment of speed and quality in the banking and finance domain. *arXiv* **2019**, arXiv:1906.01685.
37. Grangier, D.; Iter, D. The trade-offs of domain adaptation for neural language models. *arXiv* **2021**, arXiv:2109.10274.
38. Poncelas, A.; Shterionov, D.; Way, A.; Wenniger, G.; Passban, P. Investigating Backtranslation in Neural Machine Translation. *arXiv* **2018**, arXiv:1804.06189.
39. Chu, C.; Dabre, R.; Kurohashi, S. An empirical comparison of domain adaptation methods for neural machine translation. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Toronto, ON, Canada, 9–14 July 2017; pp. 385–391.
40. Deng, Y.; Yu, H.; Yu, H.; Duan, X.; Luo, W. Factorized transformer for multi-domain neural machine translation. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16–20 November 2020; pp. 4221–4230.
41. Bakr, H.A.; Shaalan, K.; Ziedan, I. A hybrid approach for converting written Egyptian colloquial dialect into diacritized Arabic. In Proceedings of the 6th International Conference on Informatics and Systems, infos2008, Cairo University, Citeseer, Cairo, Egypt, 27–28 March 2008.
42. Mohamed, E.; Mohit, B.; Oflazer, K. Transforming standard Arabic to colloquial Arabic. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Jeju Island, Republic Korea, 8–14 July 2012; pp. 176–180.
43. Habash, N.; Hu, J. Improving Arabic-Chinese statistical machine translation using English as pivot language. In Proceedings of the Fourth Workshop on Statistical Machine Translation, Athens, Greece, 30–31 March 2009; pp. 173–181.
44. Nagoudi, E.M.B.; Elmadany, A.; Abdul-Mageed, M. AraT5: Text-to-text transformers for Arabic language generation. *arXiv* **2021**, arXiv:2109.12068.
45. Zakraoui, J.; Saleh, M.; Al-Maadeed, S.; Alja'am, J.M. Arabic Machine Translation: A Survey With Challenges and Future Directions. *IEEE Access* **2021**, *9*, 161445–161468. [CrossRef]
46. Ameur, M.S.H.; Meziane, F.; Guessoum, A. Arabic machine translation: A survey of the latest trends and challenges. *Comput. Sci. Rev.* **2020**, *38*, 100305. [CrossRef]
47. Habash, N.Y. Introduction to Arabic natural language processing. *Synth. Lect. Hum. Lang. Technol.* **2010**, *3*, 1–187.
48. Ehab, R.; Amer, E.; Gadallah, M. English-Arabic hybrid machine translation system using EBMT and translation memory. *Int. J. Adv. Comput. Sci. Appl.* **2019**, *10*, 195–203. [CrossRef]

49. Sajjad, H.; Abdelali, A.; Durrani, N.; Dalvi, F. AraBench: Benchmarking Dialectal Arabic-English Machine Translation. In Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain, 8–13 December 2020; Scott, D., Bel, N., Zong, C., Eds.; International Committee on Computational Linguistics: New York, NY, USA, 2020; pp. 5094–5107. [CrossRef]
50. Khondaker, M.T.I.; Waheed, A.; Nagoudi, E.M.B.; Abdul-Mageed, M. GPTAraEval: A Comprehensive Evaluation of ChatGPT on Arabic NLP. *arXiv* **2023**, arXiv:2305.14976.
51. Antoun, W.; Baly, F.; Hajj, H. AraGPT2: Pre-Trained Transformer for Arabic Language Generation. In Proceedings of the Sixth Arabic Natural Language Processing Workshop, Kyiv, Ukraine, 19 April 2021; pp. 196–207.
52. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language Models are Unsupervised Multitask Learners. *Openai Blog* **2019**, *1*, 9.
53. Xue, L.; Constant, N.; Roberts, A.; Kale, M.; Al-Rfou, R.; Siddhant, A.; Barua, A.; Raffel, C. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online, 7 June 2021; Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., Zhou, Y., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 483–498. [CrossRef]
54. Tiedemann, J.; Thottingal, S. OPUS-MT—Building open translation services for the World. In Proceedings of the European Association for Machine Translation Conferences/Workshops, Lisbon, Portugal, 3–5 November 2020.
55. Tiedemann, J. The Tatoeba Translation Challenge—Realistic Data Sets for Low Resource and Multilingual MT. In Proceedings of the Fifth Conference on Machine Translation, Lisbon, Portugal, 3–5 November 2020; pp. 1174–1182.
56. Kudo, T.; Richardson, J. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018: System Demonstrations, Brussels, Belgium, 31 October–4 November 2018; pp. 66–71. [CrossRef]
57. Team, N.; Costa-jussà, M.R.; Cross, J.; Çelebi, O.; Elbayad, M.; Heafield, K.; Heffernan, K.; Kalbassi, E.; Lam, J.; Licht, D.; et al. No Language Left Behind: Scaling Human-Centered Machine Translation. *arXiv* **2022**, arXiv:2207.04672.
58. Popović, M. chrF: Character n-gram F-score for automatic MT evaluation. In Proceedings of the Tenth Workshop on Statistical Machine Translation, Lisbon, Portugal, 17–18 September 2015; pp. 392–395. [CrossRef]
59. Snover, M.; Dorr, B.; Schwartz, R.; Micciulla, L.; Makhoul, J. A study of translation edit rate with targeted human annotation. In Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers, Cambridge, MA, USA, 5–8 October 2006; pp. 223–231.
60. Rei, R.; Stewart, C.; Farinha, A.C.; Lavie, A. COMET: A Neural Framework for MT Evaluation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020; pp. 2685–2702. [CrossRef]
61. Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K.Q.; Artzi, Y. BERTScore: Evaluating Text Generation with BERT. *arXiv* **2020**, arXiv:1904.09675.
62. Peng, K.; Ding, L.; Zhong, Q.; Shen, L.; Liu, X.; Zhang, M.; Ouyang, Y.; Tao, D. Towards Making the Most of ChatGPT for Machine Translation. *arXiv* **2023**, arXiv:2303.13780.
63. Khoshafah, F. ChatGPT for Arabic-English Translation: Evaluating the Accuracy. 2023. Available online: https://www.researchsquare.com/article/rs-2814154/v2 (accessed on 12 July 2024).
64. Alyafeai, Z.; Alshaibani, M.S.; AlKhamissi, B.; Luqman, H.; Alareqi, E.; Fadel, A. Taqyim: Evaluating Arabic NLP Tasks Using ChatGPT Models. *arXiv* **2023**, arXiv:2306.16322.
65. Banimelhem, O.; Amayreh, W. Is ChatGPT a Good English to Arabic Machine Translation Tool? In Proceedings of the 2023 14th International Conference on Information and Communication Systems (ICICS), Irbid, Jordan, 21–23 November 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 1–6.
66. Hendy, A.; Abdelrehim, M.; Sharaf, A.; Raunak, V.; Gabr, M.; Matsushita, H.; Kim, Y.J.; Afify, M.; Awadalla, H.H. How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation. *arXiv* **2023**, arXiv:2302.09210.
67. Coughlin, D. Correlating automated and human assessments of machine translation quality. In Proceedings of the Machine Translation Summit IX: Papers, New Orleans, LA, USA, 23–27 September 2003.
68. Rei, R.; Stewart, C.; Farinha, A.C.; Lavie, A. Unbabel's Participation in the WMT20 Metrics Shared Task. In Proceedings of the Fifth Conference on Machine Translation, Online, 30 May–12 June 2020; Barrault, L., Bojar, O., Bougares, F., Chatterjee, R., Costa-jussà, M.R., Federmann, C., Fishel, M., Fraser, A., Graham, Y., Guzman, P., et al., Eds.; Association for Computational Linguistics: Portland, OR, USA, 2020; pp. 911–920.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.