Transformer Meets External Context: A Novel Approach to Enhance Neural Machine Translation

Mohammed Alsuhaibani[1], Kamel Gaanoun[2], Ali Alsohaibani[3]
Department of Computer Science, College of Computer, Qassim University, Buraydah, Saudi Arabia[1]
National Institute of Statistics and Applied Economics, Rabat, Morocco[2]
Department of English and Translation, College of Arabic Language and Social Studies,
Qassim University, Buraydah, Saudi Arabia[3]

***Abstract*—Most neural machine translation (NMT) systems rely on parallel data, comprising text in the source language and its corresponding translation in the target language. While it's acknowledged that context enhances NMT models, this work proposes a novel approach by incorporating external context, specifically explanations of source text meanings, akin to how human translators leverage context for comprehension. The suggested methodology innovatively addresses the challenge of incorporating lengthy contextual information into NMT systems. By employing state-of-the-art transformer-based models, external context is integrated, thereby enriching the translation process. A key aspect of the approach lies in the utilization of diverse text summarization techniques, strategically employed to efficiently distill extensive contextual details into the NMT framework. This novel solution not only overcomes the obstacle posed by lengthy context but also enhances the translation quality, marking a advancement in the field of NMT. Furthermore, the data-centric approach ensures robustness and effectiveness, yielding improvements in translation quality, as evidenced by a considerable boost in BLEU score points ranging from 0.46 to 1.87 over baseline models. Additionally, we make our dataset publicly available, facilitating further research in this domain.**

Keywords*—Deep learning; transformers; context; NMT; neural machine translation; natural language processing systems

I. Introduction

The rapid development of Neural Machine Translation (NMT) has changed the field of NLP, significantly improving the quality and accuracy of translations across various language pairs [1]. This advancement has facilitated cross-cultural communication, promoted cultural and intellectual understanding, and contributed to the growth of research. Despite these successes, the translation of complex source texts remains a challenge, particularly in languages that have rich morphology and complex grammar [2]. One such language is Arabic, which is characterized by a diverse range of dialects, idiomatic expressions, complex linguistic structures, rich morphology and complex grammar [3]. In order to overcome these challenges, it is essential to develop NMT systems that are capable of incorporating contextual information to produce more accurate translations that closely resemble human-like understanding and interpretation.

Drawing inspiration from the practices of human translators and language professionals, this study aims to develop a context-aware NMT model for the Arabic language that moves beyond existing translation approaches. In particular, this work is built upon the concept of "deverbalization," as proposed by [4] and further elaborated by [5], which emphasizes the importance of comprehending the underlying meanings conveyed in the source language rather than merely interpreting linguistic symbols. By focusing on contextual details, idiomatic expressions, and cultural subtleties, this approach has the potential to capture the essence of the source text more accurately, leading to more faithful translations that are true to the intended meaning of the original Arabic text.

This research explores the use of explanatory data as a source of contextual information to be injected into the NMT model. The explanatory data, derived from authoritative sources such as Quran exegesis books, provides valuable insights into the intended meaning and cultural context of the Arabic text. By incorporating this context-rich data into the model, the aim is to enable the NMT system to better comprehend the source text, capture the linguistic subtleties, and produce translations that are more accurate and faithful to the original text.

To the best of the authors knowledge, neither previous research has investigated the impact of contextual information, nor has any work explored the use of "explanations" of Arabic text as context information on Arabic NMT. Furthermore, there is a lack of Arabic corpora that include texts accompanied by their explanations. This study seeks to fill this gap by proposing a new dataset and a context-aware NMT model for the Arabic language. By doing so, it is intend to contribute to the body of knowledge on context-aware NMT systems and highlight the significance of incorporating contextual information in improving translation accuracy.

In this research, the state-of-the-art T5 (Text-to-Text Transfer Transformer) [6] model, in its multilingual version (mT5) [7], is employed. This model has demonstrated exceptional performance in various natural language processing tasks, including machine translation [8]. The T5 model, characterized by its advanced architecture and robust pre-training capabilities, serves as an optimal foundation for the context-aware NMT model within the framework of this work. By fine-tuning the mT5 model on the context-rich dataset, the assessment is conducted on the model's ability to leverage contextual information to enhance translation accuracy and deliver translations that are more faithful to the original Arabic text. Through this research, the aim is to contribute to the ongoing efforts to improve the quality and accuracy of Arabic NMT systems by incorporating contextual information, paving the way for more reliable and human-like translations that respect the linguistic details and richness of the Arabic language. Moreover, the findings are anticipated to yield benefits not solely for the Arabic NMT community but also to inspire researchers engaged in languages with akin complexities. This may prompt exploration into the potential of context-aware NMT models, consequently augmenting the overall performance and capabilities of NMT systems across diverse language pairs.

This study stands out by offering four main contributions to the body of knowledge:

- Introducing a novel strategy to improve NMT by incorporating contextual data extracted from source-specific explanatory materials. This method aims to enhance the translation's precision and intelligibility, regardless of the original language or subject matter.
- Fine-tuning a multilingual T5 model (mT5) using two newly proposed datasets, one that pairs source content with its corresponding translations, and another that enriches the source content by integrating it with relevant contextual details.
- Proposing three different methods of context injection: one utilizing the complete explanatory content, another employing a summarized version of this content, and the third adding an additional identifying detail along with the summarized explanation to the original content. The novel yet simple approaches, effectively simplify the handling of long context in NMT.
- Curating and disseminating an innovative Arabic-English parallel dataset, enriched with comprehensive explanations of the source text, thereby offering a robust resource for advancing machine translation research and facilitating the exploration of contextual augmentation in NMT systems.

The structure of this paper unfolds as follows: in Section II, a comprehensive examination of pertinent literature is explored, elucidating the current state of NMT research and contextual information, while shedding light on the existing gaps our study aims to bridge. Section III describes the detailed process of data collection and preprocessing, resulting in the creation of the parallel dataset, which encompasses original Arabic verses, their respective explanations, and corresponding English translations. In Section IV, the proposed context-aware NMT model is described, elaborating on the methodology employed to integrate contextual information. Section V explains the experimental framework, encompassing the fine-tuning of the T5 model, the evaluation metrics, and the results obtained from the experiments alongside a detailed analysis of a qualitative nature. Section VI goes into a thorough discussion, exploring various perspectives and interpretations of the findings. Finally, Section VII summarizes the research conclusions, highlighting the implications of the findings and suggesting potential paths for future research in context-aware NMT.

II. Related Work

Several works have been conducted to explore such potential benefits of integrating contextual data into NMT systems. For example, [9] proposed a study that centres on augmenting the NMT architecture by leveraging the surrounding text as an essential source of contextual information. To achieve this goal, the researchers extended the attention-based NMT method by introducing an additional set of an encoder and attention model that encodes the context sentence immediately preceding the current source sentence. Experimental evaluations were performed on the English-French and English-German language pairs, where the obtained results demonstrated that the proposed methodology significantly outperformed the baseline models that did not incorporate the surrounding text. Similarly, [10] aimed to improve the quality of NMT by introducing a hierarchical attention model that captures the context in a structured and dynamic manner. The study conducted experiments on Chinese-to-English (Zh-En) and Spanish-to-English (Es-En) datasets, and the proposed model was integrated into the original NMT architecture as an additional level of abstraction, conditioned on the NMT model's previous hidden states. The experimental results showed that the hierarchical attention model significantly outperformed the baseline models in terms of translation quality and fluency, demonstrating the effectiveness of incorporating context in a structured and dynamic manner.

The author in [11] investigated the incorporation of contextual information in NMT by modifying the transformer [12] model for context-agnostic NMT to handle additional context. The experiments were conducted on an English-Russian subtitles dataset, and the modified model first encoded the source sentence and the context sentence independently. Then, a single attention layer, in combination with a gating function, was utilized to generate a context-aware representation of the source sentence. The results showed that the proposed model outperformed the baseline models in terms of translation quality and fluency, highlighting the importance of incorporating contextual information in NMT. Moreover, [13] addressed the core challenge of effectively encoding and aggregating contextual information and proposed a novel approach that utilized a pre-trained BERT [14] model as an additional encoder to encode contextual information with German to English and vice versa. This resulted in a group of features carrying contextual information that was subsequently incorporated into the attention mechanism of the NMT model. The experiments showed that the proposed approach improved translation quality and fluency. These findings highlight the importance of contextual information in NMT and demonstrate the potential of leveraging pre-trained models to effectively incorporate contextual cues into the translation process.

Furthermore, [15] focused on exploring the ability of NMT to discover cross-sentential dependencies in the absence of explicit annotation or guidance. Specifically, the study examined the translation of movie subtitles from German to English and aimed to identify the impact of additional contextual information on the translation and attention mechanisms of the NMT model. Unlike prior research that modified the NMT model by adding a separate context encoder and attention mechanism, the study modified only the input and output segments while keeping the standard setup. The research team conducted a series of experiments with different context windows and evaluated two models that extended context in different ways: extended source and extended translation units. The former included context from the previous sentences in the source language to improve the encoder part of the network, while the latter involved translating larger segments of the source language into corresponding units in the target language. The findings suggest that incorporating additional contextual information in the NMT model can improve the quality and fluency of translations and highlight the importance of further research in this area.

On the other hand, one prominent direction in NMT is the exploration of data-centric approaches for enhancing NMT with contexts. These approaches prioritize leveraging the available multilingual data to boost translation quality, instead of solely focusing on model architecture or algorithmic improvements.

For instance, [16] introduced an approach to enhance NMT by incorporating the entire document context. This method involves pre-processing to add contextual information from each document to its respective sentences, thereby aiming to improve translation coherence and resolve cross-sentential ambiguities. They propose using a simple method to estimate document embeddings, involving averaging all word vectors in a document to maintain a consistent dimension. The technique is applied to a Transformer base model and tested on English-German, English-French, and French-English language pairs. Similarly, [17] explored the impact of incorporating extended context into attention-based NMT, particularly focusing on translated movie subtitles. The study opted to adjust the input and output segments while retaining the standard model setup. It primarily investigates the capacity of NMT to recognize cross-sentential dependencies without specific annotations or guidance.

The reviewed literature demonstrates the benefits of incorporating contextual information from the source language in NMT models for various language pairs. However, to the best of authors knowledge, no previous work has investigated the effect of contextual information on Arabic NMT. Moreover, neither previously proposed work explored the use of “explanations” of Arabic text as context information nor does Arabic corpora exist that include texts accompanied with their explanation. Therefore, this study aims to fill this gap by proposing a new dataset and a context-aware NMT model for Arabic.

III. Data

In order to develop a context-aware Arabic NMT model, it is essential to acquire parallel data encompassing Arabic text, its corresponding English translation, and explanatory information for the Arabic text. This section elucidates the data sources, outlines the criteria guiding their selection, and details the procedures for data collection and preprocessing that were applied to generate the final dataset.

The criteria for selecting the data sources are as follows:

- Presence of parallel (Arabic - English) data
- Availability of explanations for the Arabic text
- Sequential numbering of text segments in both Arabic and English versions to facilitate alignment
- Numbering or clear linking of explanations to their corresponding text
- Clear indication of explanations to facilitate the scrapping process

Before going into the specifics of the data sources, it is crucial to provide a concise overview of “explanations” within the context of Arabic literature. Explanations, frequently encountered in Arabic scholarly works, serve as commentaries or interpretations that illuminate the meaning, context, and significance of the original text. They play a vital role in clarifying the intended meaning, exposing linguistic subtleties, and offering historical and cultural insights. Consequently, they contribute to enhancing the reader’s comprehension of the text.

For this study, the Saadi Exegesis was utilized as a primary data source, as accessed through the Islamweb website. Selected for its simplicity and brevity compared to other exegesis sources, the Saadi Exegesis provides concise explanations for each verse in the Quran, with numbered explanations corresponding to verse numbers. The scraping process involved acquiring explanations for each of the 114 Surahs, manually correcting verse numbers when necessary, and ultimately compiling a corpus of 37 Surahs with corresponding verse numbers. Subsequently, the Saadi corpus was merged with a parallel Arabic-English Quran dataset. The parallel Arabic-English Quran dataset was created by scraping the English translations of the Quran[1], which offers seven reputable English translations. Among these translations, Sahih International was chosen for its widespread recognition as a popular translation, renowned for employing a communicative translation approach that prioritizes both linguistic clarity and accuracy. Additionally, as noted by [18], it focuses on simplifying and clarifying English, deliberately avoiding unnecessary transliterations to ensure broad accessibility.

The Arabic verses were retrieved from the Mendeley repository[2]. To guarantee compatibility between the Saadi and parallel datasets, the Surah names in the Saadi dataset underwent preprocessing, involving the removal of hamza letter ”ء“ and the renaming of certain Surahs. The ultimate merged dataset encompasses 2,908 verses from 37 Surahs, with the following fields: Surah number, Verse number, Verse text, Translation, and Tafseer (exegesis).

In summary, the comprehensive efforts in data collection and preprocessing have led to the creation of a novel dataset that combines parallel source-target sequences with corresponding source explanations. This dataset, encompassing parallel Arabic-English texts and contextual information in the form of explanations, establishes a robust foundation for the development and evaluation of the context-aware Arabic NMT model. Table I presents the statistics and details of the dataset. It is important to emphasize that the dataset is publicly available[3], with the aim of facilitating and encouraging future research by providing an accessible resource for the scientific community. The subsequent sections of this paper will elaborate on the proposed model and the methodology employed to incorporate contextual information, along with the experimental setup and the results obtained from the experiments.

[1] https://corpus.quran.com/
[2] https://data.mendeley.com/datasets/sg5j65kgdf/2
[3] https://github.com/KamelGaanoun/CAANMT

TABLE I. KEY STATISTICS OF THE PROPOSED DATASET

Language	Verses			Explanations		
	Total	Average length	Median length	Total	Average length	Median length
Arabic	2908	16	14	2908	117	87
English		33	29	-	-	-

IV. PROPOSED APPROACH

The objective is to develop a context-aware Arabic NMT model aimed at improving the translation quality of Arabic text. To achieve this, the multilingual T5 (mT5) model is utilized, and transfer learning is applied to the domain and data of the study. A data-centric approach is also adopted by incorporating context as a data augmentation into the source language. In the following subsections, the T5 model is introduced, followed by a description of the global process.

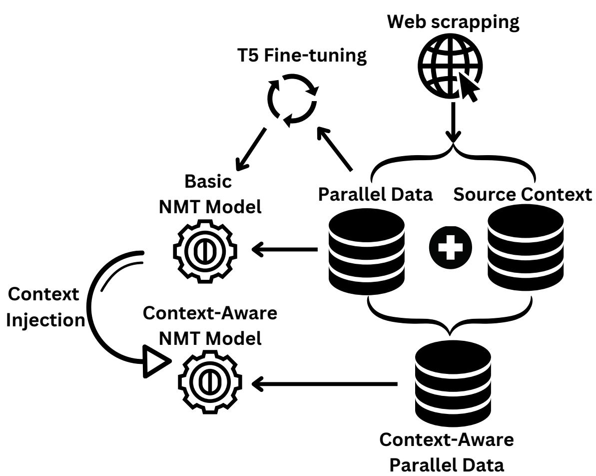

Fig. 1. An overview of the proposed approach global process and architecture.

A. T5 Model

The foundation of the context-aware Arabic NMT approach proposed in this study rests on the Transformer-based T5 model architecture. Initially introduced by Raffel et al. in 2020 [6], T5 adopts the Transformer encoder-decoder structure and undergoes pre-training on a substantial text corpus in a self-supervised manner. T5 models treat input text as a sequence of tokens and generate output text, making them particularly suitable for text-to-text tasks. By unifying all NLP problems into a text-to-text format, T5 establishes a cohesive framework for transfer learning across various tasks. In the context of translation, the source sentence becomes the input text, and the translated sentence is the target.

The encoder utilizes self-attention and feedforward layers to map input tokens into a contextual representation. This context-rich encoding is then transmitted to the decoder, which generates the output text token by token using cross-attention, relying on the encoder output. T5 incorporates relative position embeddings to denote positional information between tokens. Fig. 2 shows a high-level overview of the T5 architecture.

Through pre-training on extensive datasets, T5 models acquire universal text representations, facilitating generalization across downstream tasks via transfer learning. During pre-training, the objective is to predict randomly masked spans of input text, akin to the approach in BERT [14].

This study employs the multilingual T5 variant (mT5) [7], pre-trained on 101 languages, including Arabic. mT5 has demonstrated state-of-the-art performance on translation benchmarks, surpassing previous models. By fine-tuning mT5 on the proposed context-aware Arabic dataset, the model effectively leverages explanatory details to enhance its translation capabilities. The transformer architecture's ability to model cross-sentence context proves critical for this task. In summary, T5 establishes a fitting foundation for the development of the context-aware NMT system in this research.

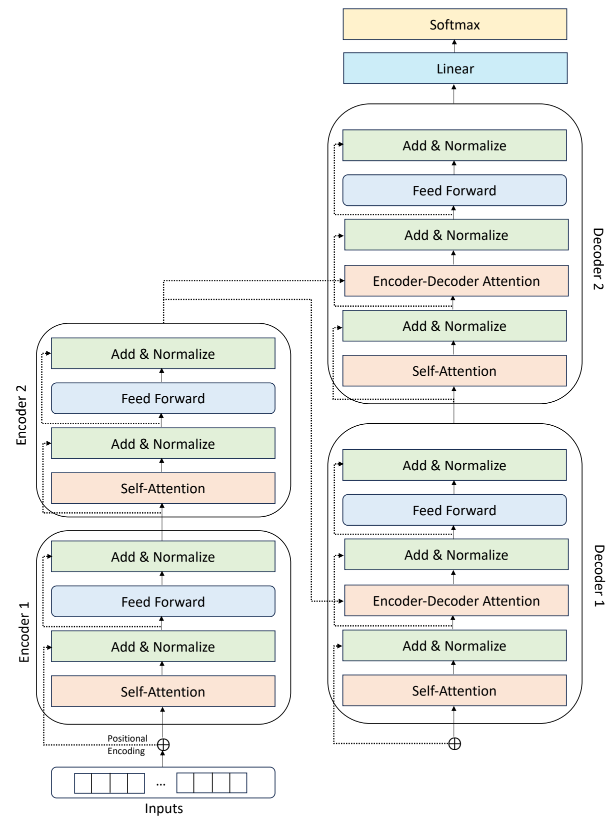

Fig. 2. T5 simplified architecture.

B. Context Aware NMT process

The methodology for developing a context-aware Arabic NMT model unfolds through a systematic, four-step process. Initially, the crucial task of data collection involves assembling a parallel dataset comprising original Arabic verses, their corresponding explanations from a Quran exegesis book, and their respective English translations. Subsequently, the context injection phase ensues, wherein verses are adroitly concatenated with their pertinent contextual information. The third step entails fine-tuning a T5 model on each dataset individually; first on the dataset containing verses and their translations, and subsequently on the context-aware dataset with the injected context. Finally, the results gleaned from

TABLE II. DATASET PARTITIONS STATISTICS

Partition	Number of verses
Training	2000
Development	800
Test	108

both datasets are scrutinized, drawing insightful conclusions regarding the efficacy of the context injection approach and its impact on the performance of the NMT model. Fig. 1 summarizes the proposed approach and its global process.

V. EXPERIMENTS AND RESULTS

A. Experimental Settings

The dataset was meticulously partitioned into three distinct subsets: a training set comprising 2000 verses, a development set encompassing 800 verses, and a test set consisting of 108 verses (Table II). The configuration used for the experiments includes an NVIDIA A100-SXM 40GB GPU, Python 3.9, and SimpleTransformers 0.63.9 library. The main set of training parameters are presented in Table III.

Throughout the experiments, transfer learning is employed by fine-tuning a T5 model on the corresponding training data. The challenge arises from the limited availability of T5 models specifically tailored for Arabic text. The AraT5 model, despite being designed for Arabic, proved suboptimal for translation tasks. Consequently, the mT5 model—a multilingual T5 variant proficient in Arabic and 100 other languages—emerges as the only viable alternative. Due to time and computing resource constraints, the experiments utilize the small version of mT5 (mT5-small).

To ensure comparability and reproducibility of results, all experiments on both the Baseline and context-aware datasets are conducted with identical parameters and a fixed seed. Initially, the models undergo a training process of 5 epochs to establish a preliminary benchmark for comparison purposes and to determine the optimal sequence length and batch size. Subsequently, models exhibiting superior performance relative to the Baseline model are selected for further training over a period of 15 epochs.

TABLE III. MAIN TRAINING PARAMETERS

Parameter	Value
Learning Rate	0.001
Manual Seed	2023
Warmup Ratio	0.06
Adam Betas	(0.9, 0.999)
Adam Epsilon	$1e^{-8}$

Context injection is conducted through various methods:

- Method 1: Leveraging the entire verse explanation as contextual information.
- Method 2: Employing a summarized version of the explanation, achieved through:
    - *KeyBERT option 1*: Extracting keyphrases from the explanation text using the KeyBERT [19] library. Diverse keyphrases are generated by activating the Max Sum Distance parameter. For instance, three sets of keyphrases, each containing four words, are concatenated for every verse explanation.
    - *KeyBERT option 2*: Alternatively, dividing the explanation of each verse into three segments, from which two distinct keyphrases consisting of three words are generated and concatenated.
    - Abstractive summaries using three Transformer-based encoder-decoder models designed for Arabic text summarization. Two of these models utilize the T5 architecture [20], [21], and the third employs mBert in both the encoder and decoder modules [20]. Similar to KeyBERT option 2, the models are applied to the three segments of each verse's tafssir.
- Method 3: Concatenating the Surah name to the Verse, in addition to the explanation injected as in Method 2.

Addressing the long context problem outlined by [22], where two solutions are proposed, this work's novel approach diverges by being data-centric and relying on summarization, offering a distinct methodology in dealing with long contextual information. Indeed, the use of the entire context is not possible due to the limited number of tokens supported by the models, not to mention the difficulty encountered by the model in finding the most relevant parts to consider for translation in a long context. By attempting to integrate the entire context, the model only uses the beginning of the text, and thus loses most of the information. The proposed method has the advantage of not requiring any changes to the model structure, and relies solely on the way in which the context is integrated.

Methods 2 and 3 are specifically designed to address the challenge of lengthy explanations overwhelming the models and causing attention to be scattered across the entire text. These techniques aim to concentrate the explanation by highlighting critical elements or providing a summary of the text. This approach mirrors human information processing, selectively retaining essential components to comprehend the overarching concept and enhance the translation process.

Both extractive and abstractive summarization techniques are employed to extract the most pertinent contextual elements. The extractive method aims to retrieve keyphrases verbatim from the context text, whereas the abstractive approach utilizes alternative wording akin to human summarization techniques, employing paraphrasing and synonyms.

The concatenation process employs designated tokens, represented as EXP for explanation and CHAP for the Surah name, as illustrated in Table IV. To prevent potential confusion with explanation words, these tokens are expressed in Latin characters. Moreover, they are deliberately chosen to be concise and correspond to a single token following T5 tokenization.

Various experiments were compared based on the tested parameters and configurations. Specifically, two batch sizes

TABLE IV. CONCATENATION METHODS

Method	Concatenation
1	Verse explanation + EXP + Verse
2	Summarized Verse explanation + EXP + Verse
3	Summarized Verse explanation + EXP + Surah name + CHAP + Verse

were evaluated: 8 and 16. Additionally, the maximum input length varied between 200 and 400, incorporating different methods of context injection and considering whether or not to employ a preprocessing step for explanations.
The nomenclature for each model is defined based on the combination of adopted parameters. For instance, a model using a batch size of 8, a maximum length of 200, and including context will be named EXP200bs8. A similar model but incorporating a summarized version of the context will be named EXP200bs8XX, where XX represents the name of the summarization model used. In addition to this nomenclature, a number is assigned to each model to facilitate interpretation and comparisons. The various configurations tested are detailed in the Results section(V-C) and Table V.

B. Evaluation Metric

To appraise the influence of context injection on neural machine translation performance, the widely recognized BLEU (Bilingual Evaluation Understudy) [23] metric was employed. As an automatic evaluation metric, the BLEU metric is extensively utilized for assessing machine translation output. It quantifies the similarity between the predicted translation and one or more reference translations based on the n-gram overlap between them.

The BLEU score is computed using the following equation:

$$BLEU = BP \times exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

where BP represents the brevity penalty, N denotes the maximum order of n-grams considered, w_n signifies the weight assigned to n-grams of order n, and p_n corresponds to the precision of the predicted n-grams in the reference translations of order n.

The precision of the predicted n-grams (up to a certain order) in the reference translations is calculated and subsequently combined using a geometric mean. In the conducted experiments, the BLEU score for each translation model is reported under varying experimental conditions, such as with or without context injection, or with different types of context injected. The enhancement in BLEU score relative to a baseline model devoid of any context is also reported. By employing the BLEU metric and its equation, a quantitative and objective evaluation of the impact of context injection on NMT performance is provided. This enables a comparison of diverse context injection techniques and the drawing of meaningful conclusions regarding their effectiveness.

The current study utilizes SacreBleu 2.3.1, a library introduced by [24], for the computation of BLEU scores. The adoption of this library hinges on its ability to ensure comparability and reproducibility of evaluations. As a result, future research may employ the default parameters of this library to facilitate comparisons between different systems.

C. Results

TABLE V. TEST SET BLEU SCORES FOR DIFFERENT MODELS. THE NAME OF EACH MODEL IS BASED ON ITS PARAMETERS (EXP: MODEL INCLUDES EXPLANATIONS; BS: BATCH SIZE; PRC: EXPLANATION IS PREPROCESSED; KEY1,KEY2,ETC: USED SUMMARIZATION MODEL; 200,400,ETC: MAXIMUM SEQUENCE LENGTH). A NUMBER IS ALSO ASSIGNED FOR EASY REFERENCE.

Model	5 Epochs	15 Epochs	Enhancement
Baselines			
(1) 200bs16	3.95	-	-
(2) 400bs8 (Baseline)	5.27	23.07	-
Context-Aware models			
No summarization			
(3) EXP200bs16	2.70	-	-
(4) EXP200bs16PRC	3.20	-	-
(5) EXP300bs16PRC	3.14	-	-
(6) EXP400bs8PRC	4.17	-	-
With summarization			
(7) EXP400bs8PRCKey1	6.11	24.85	7.7% (+1.78)
(8) EXP400bs8AraT5Titles	2.91	-	-
(9) EXP400bs8AraT5Sum	2.53	-	-
(10) EXP400bs8mBert2mBert	6.91	23.53	2% (+0.46)
(11) EXP400bs8mBert2mBertCHAP	6.97	24.94	8.1% (+1.87)
(12) EXP400bs8Key2	5.72	23.56	2.1% (+0.49)
(13) EXP400bs8Key2CHAP	7.68	23.60	2.3% (+0.53)

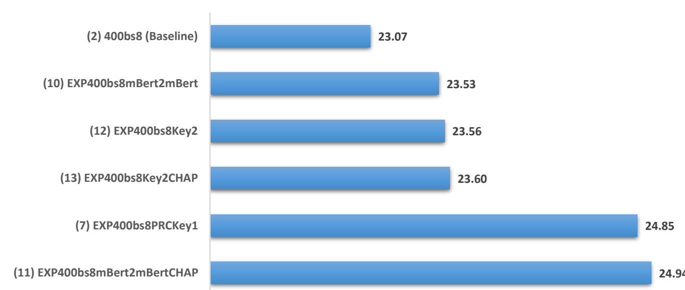

Fig. 3. Model's performance comparison with 15 epochs.

Table V presents the final results for different configurations. In the initial experiments, a first baseline model (1) was trained with a batch size of 16 and a sequence length of 200 for 5 epochs. This model, consisting of the original input and translation, attained a BLEU score of 3.95. Following this, context-aware models were examined using the same configuration. Although these models demonstrated lower performance compared to the baseline, preprocessing the explanation text led to an improvement in their performance. As a result of this preprocessing, the BLEU score rose from 2.7 (3) to 3.2 (4). The preprocessing included some letter normalization, as illustrated in Table VI, and stopwords suppression. The list of stopwords is derived from the NLTK library and augmented by a list of Quran explanation-related stopwords gathered by the authors, this list is also publicly available[4].

Moreover, increasing the sequence length to 400 improved the score by one point (6). Subsequent experiments were conducted using the same configuration, involving a sequence

[4]https://github.com/KamelGaanoun/CAANMT

TABLE VI. NORMALIZED LETTERS DURING PREPROCESSING PROCESS

Original form	**Normalized form**	**Rule**
إ، أ، آ، ٱ، ا	ا	Hamzah (ء) removal
ى	ي	Final Yaa(ي) rule
ؤ، ئ	ء	Hamzah simplification
ة	ه	Taa Marbutah (ة) normalization
لـــ	ل	Elongation removal

length of 400 tokens and a batch size of 8. The second baseline model with this configuration (2) achieved a score of 5.27, exceeding the performance of the first context-aware models by over one point. Model (2) will be retained as the final Baseline for all following experiments.

All subsequent models were constructed using a summarized version of the explanatory text. With the exception of models (8) and (9), which employed AraT5 summarizing models, all other models exhibited superior performance compared to the baseline (2). These models were selected for additional training over ten epochs to validate the initial findings. Indeed, they outperformed the baseline model by a margin ranging from 0.46 to 1.87 BLEU score points. The most significant improvement was achieved by model (11), based on mBERT in both encoder and decoder modules for summarization, and incorporating the surah name as a global context. This model attained a remarkable overall score of 24.94. Fig. 3 renders the results, imbuing them with heightened readability.

D. Qualitative Analysis

As demonstrated in Table VII, examples 1 and 2 conspicuously illustrate the proficiency of the proposed model, which draws upon targeted explanations of specific components within a verse to enhance translation accuracy. The initial example showcases the limitation of the baseline model, as it misses the term "Woe" in its translation. In contrast, the best model astutely renders the term ويل as "Woe", reflecting the contextual guidance provided. Similarly, the second example underscores the translation of the word بلغ, which the baseline model neglects, yet the best model renders correctly as "notification", adhering strictly to the original translation. It is noteworthy that the explanation for Example 2 utilizes a synonym, تحذير, to shed light on the term, while the best model leverages the original translation word, demonstrating its adeptness at comprehending context.

Example 3 presents a discernible contrast between the translations of the baseline and best models. In this instance, the best model generates a translation that, while not entirely verbatim, proves more effective compared to the baseline model, which unfortunately projects a meaning counter to the original text's intent. This best performance of the advanced model can be ascribed to its capability to optimize the provided explanation for the phrase عبادنا المؤمنين signifying "believing servants" or "believers".

The advanced model also displays an impressive aptitude to generate precise translations, even in the absence of specific lexical guidance within the explanation. This proficiency is evident in Example 4, where the best model correctly introduces the term "Deny" in its translation, while the baseline model erroneously uses "lie", a distorted representation of the intended meaning of تكذبون. The ambiguous nature of تكذبون, which may denote "Lying" or "Denying" based on the diacritical marks used, calls for judicious contextual interpretation, a strength displayed by the best model.

Despite the marginal difference in the score indicated in Example 5, the best model outperforms in creating translations that hew closely to the original text. Interestingly, the accompanying explanation does not proffer any clear insight into the specific translation process, yet the model maintains accuracy.

The detailed analysis suggests that the best model reaps benefits not solely from the immediate verse explanation but also from its past experience with other verses' explanations. Moreover, the model benefits from incorporating the Sourah name as a global contextual cue. This advantage is evident when comparing the Verse translation in Example 4 produced by the best model with a version that excludes the use of the Sourah name in its context. In this case, the latter model inaccurately uses "lie" instead of "deny".

An important observation pertains to the robustness of the translation models towards out-of-context summaries. Typically, summarisation models are trained on news and social media corpora and hence may lack proficiency in classical Arabic or Arabic literature. This could lead to some out-of-Quranic-context summaries in the present work. Yet, the translation models perform consistently on the input texts and resist confusion induced by these out-of-context segments, as demonstrated across all examples.

VI. DISCUSSION

The experimental results demonstrate the effectiveness of the context-aware neural machine translation model in improving translation accuracy for the Arabic language. This section discusses the key findings and implications of the study, addresses limitations, and explores potential future directions.

Incorporating contextual information, specifically verse explanations derived from Arabic book exegesis, significantly enhances translation quality. The context-aware models achieved BLEU scores ranging from 23.53 to 24.94, representing an improvement of up to 1.87 points or 8.1% in relative BLEU score compared to the baseline model without context. Leveraging external sources of context enables the NMT system to capture linguistic subtleties and produce more accurate translations by providing additional insights into the intended meaning and cultural context of the source text.

Moreover, using a summarized version of the explanation text improves translation performance compared to using the complete explanation. Models utilizing summarized explanations achieved higher BLEU scores, indicating that concentrating the explanation by emphasizing critical elements or providing a summary enhances translation quality. This finding aligns with human information processing, where selective retention of pertinent components aids understanding. As noted by [25]: "*...explained this concept to describe an independent stage*

TABLE VII. QUALITATIVE ANALYSIS

Original Verse	Context-Aware Input	Original Translation	Baseline Translation	Best Model Translation	Score difference
وقالوا يٰويلنا هٰذا يوم الدين	نصائح لمواجهة الزهور للويل والثبور دعوات ديون الدنمارك تهدد CHAPالصافات EXPبفقراء وأقرانهم وقالوا يٰويلنا هٰذا يوم الدين	They will say, "O woe to us! This is the Day of Recompense."	And they say, "O our father, this is our Day of Recompense."	And they say, "O woe to us, this is our Day of Recompense."	19.3
وما علينا الا البلٰغ المبين	نصائح للحفاظ على الأمور المطلوبة تدابير لمواجهة العذاب تحذير أممي EXP من اهتكارات الكمامات وما علينا الا البلٰغ المبين CHAPيس	And we are not responsible except for clear notification."	And We do not accept the repentance of clear proofs.	And upon us is not but a clear notification.	10.7
انه من عبادنا المؤمنين	عبادنا المؤمنين ندوة الإيمان إلى درجة اليقين ملكوت السماوات CHAPالصافات EXP والأرض انه من عبادنا المؤمنين	Indeed, he was of Our believing servants.	Indeed, He is of those who associate others with Allah	Indeed, from Our servants are the believers.	3.7
هٰذا يوم الفصل الذي كنتم به تكذبون	عيد ميلاد العبارة في يوم الفصل بين العبيدي والعباس ربائن من الحقوق الشرعية نصائح للتخلص من CHAPالصافات EXPالاكتئاب الذي كنتم به تكذبون هٰذا يوم الفصل	[They will be told], "This is the Dayof Judgement which you used to deny."	That is the Day of Recompense which you used to lie.	This is the Day of Resurrection which you used to deny.	14.1
وما علمنٰه الشعر وما ينبغي له ان هو الا ذكر وقرءان مبين	محمد بن راشد يروي تفاصيل مهمة عن شعره صلاة الضالون على النبي محمد EXPالمعمول يحذير من فطر العقول وما علمنٰه الشعر وما ينبغي له انCHAPيس هو الا ذكر وقرءان مبين	And We did not give Prophet Muhammad, knowledge of poetry, nor is it befitting for him. It is not but a message and a clear Qur'an	And We have taught him writing, but it is not but a reminder and a clear Qur'an.	And We have not taught him, but it is not but a reminder and a clear Qur'an	0.31

where meaning is abstracted from the language forms... ". Various summarization techniques, including keyphrase extraction and abstractive summarization models, yielded improvements in translation accuracy.

Additionally, incorporating the Surah name as a proxy for global document context further improves translation quality. Models that concatenated the Surah name along with the summarized explanation and verse achieved better results compared to models without the Surah name. This finding suggests that higher-level context, such as the Surah name, provides additional cues for the NMT system to better understand and translate the source text, capturing religious and cultural connotations associated with specific Surahs. This finding confirms the importance of global document context, as highlighted by [26]. However, unlike previous work, the current approach employs a shortcut to obtain global context by using a general topic illustrated in the Surah name.

The study also highlights challenges and limitations of context-aware NMT for Arabic. One major challenge is the lack of dedicated datasets for this work in Arabic. Creating a custom dataset with paired source content and translations, enriched with explanatory details, addressed this issue. However, larger and more diverse datasets would contribute to the development and evaluation of context-aware NMT systems for Arabic.

Another limitation is the absence of specialized summarization models for Classical Arabic, which affects the quality of the summarized explanations. General-purpose summarization models trained on Modern Standard Arabic (MSA) or multilingual data were utilized, potentially missing the unique characteristics and nuances of Classical Arabic texts. Developing dedicated summarization models specifically for Classical Arabic could improve the effectiveness of the context-aware NMT approach. Moreover, there is a lack of specialized summarization models for literature and exegesis books in Arabic. The available summarization models are primarily designed for newspapers, blogs, and general content. This mismatch in specialization hinders the quality of the summarized explanations for the context-aware NMT. Developing dedicated summarization models specifically tailored to the unique characteristics and nuances of Classical Arabic texts, such as literature and exegesis books, would greatly enhance the effectiveness of the context-aware NMT approach.

Regarding future directions, expanding the dataset with more diverse genres and sources of explanatory content would enhance the generalization and coverage of the context-aware models. Leveraging advancements in natural language processing, such as pre-training techniques and transformer-based models, could further improve the performance of the context-aware NMT model. Fine-tuning larger and more sophisticated models, such as the full-size T5 or domain-specific models, may yield even better translation results. This method can be applied to other languages using the same process and is also extensible to fields such as historical texts or any area requiring expert advice and explanation. It can be employed whenever an external context exists, particularly when that context is long. The explanation of the source text can be substituted with any text that offers additional information about the text to be translated, such as expert advice, witness testimonies, or comments from teachers or professors.

Exploring alternative methods for injecting context, such as leveraging linguistic annotations or semantic representations, could provide further insights into the impact of different types of context.

VII. Conclusion

In conclusion, this study looked into the influence of injecting explanatory context on the performance of neural machine translation in rendering Arabic religious texts into English. Various methods of context injection were explored, ranging from using the entire verse explanation to incorporating summarized versions of the explanation or keyphrases extracted from it. The multilingual T5 (mT5-small) model was fine-tuned on different datasets, including context-aware and baseline datasets, to assess the effectiveness of the context injection techniques.

The results revealed that context-aware models generally exhibited superior performance compared to the baseline model, particularly when utilizing a summarized version of the explanatory text or incorporating the surah name as a global context. Furthermore, the findings demonstrated the importance of preprocessing the explanation text and carefully selecting the appropriate sequence length and batch size for training the models. The most notable improvement in translation quality was achieved by the model that employed the mBERT 2 mBert summarization technique and incorporated the surah name as a global context, achieving a remarkable overall BLEU score of 24.94. This practice aligns with the deverbalization concept, emphasizing the paramount importance of comprehending the underlying meanings conveyed in the source language over a mere interpretation of linguistic symbols.

These findings underscore the potential of context injection as a valuable approach to enhance NMT performance in translating Arabic religious texts, providing a more accurate and contextually rich understanding of the original text. Future research could explore other context injection techniques or expand the scope of this study to include other languages or genres of text. By refining and optimizing context injection methods, researchers can contribute to the development of more sophisticated and effective NMT systems, ultimately facilitating accurate and meaningful translation across diverse linguistic and cultural contexts.

Acknowledgment

This research received grant no. (180/2022) from the Arab Observatory for Translation (an affiliate of ALECSO), which is supported by the Literature, Publishing & Translation Commission in Saudi Arabia.

References

[1] F. Stahlberg, "Neural machine translation: A review," *Journal of Artificial Intelligence Research*, vol. 69, pp. 343–418, 2020.

[2] S. Berrichi and A. Mazroui, "Addressing limited vocabulary and long sentences constraints in english–arabic neural machine translation," *Arabian Journal for Science and Engineering*, vol. 46, no. 9, pp. 8245–8259, 2021.

[3] M. M. Mahsuli, S. Khadivi, and M. M. Homayounpour, "Lenm: Improving low-resource neural machine translation using target length modeling," *Neural Processing Letters*, pp. 1–32, 2023.

[4] D. Seleskovitch, "Language and cognition," *Language interpretation and communication*, pp. 333–341, 1978.

[5] Lederé, "The interpretation and translation theory of interpretational school," pp. 7–12, 2001.

[6] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485–5551, 2020.

[7] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel, "mt5: A massively multilingual pre-trained text-to-text transformer," *arXiv preprint arXiv:2010.11934*, 2020.

[8] J. Son and B. Kim, "Translation performance from the user's perspective of large language models and neural machine translation systems," *Information*, vol. 14, no. 10, 2023. [Online]. Available: https://www.mdpi.com/2078-2489/14/10/574

[9] S. Jean, S. Lauly, O. Firat, and K. Cho, "Does neural machine translation benefit from larger context?" *arXiv preprint arXiv:1704.05135*, 2017.

[10] L. Miculicich, D. Ram, N. Pappas, and J. Henderson, "Document-level neural machine translation with hierarchical attention networks," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 2947–2954. [Online]. Available: https://aclanthology.org/D18-1325

[11] E. Voita, P. Serdyukov, R. Sennrich, and I. Titov, "Context-aware neural machine translation learns anaphora resolution," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 1264–1274. [Online]. Available: https://aclanthology.org/P18-1117

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, . Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[13] X. Wu, Y. Xia, J. Zhu, L. Wu, S. Xie, and T. Qin, "A study of bert for context-aware neural machine translation," *Machine Learning*, vol. 111, no. 3, pp. 917–935, 2022.

[14] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.

[15] J. Tiedemann and Y. Scherrer, "Neural machine translation with extended context," in *Proceedings of the Third Workshop on Discourse in Machine Translation*. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 82–92. [Online]. Available: https://aclanthology.org/W17-4811

[16] V. Macé and C. Servan, "Using whole document context in neural machine translation," in *Proceedings of the 16th International Conference on Spoken Language Translation*, 2019.

[17] J. Tiedemann and Y. Scherrer, "Neural machine translation with extended context," in *Proceedings of the Third Workshop on Discourse in Machine Translation*, 2017, pp. 82–92.

[18] S. Qudah-Refai, "Dogmatic approaches of qur'ān translators: Linguistic and theological issues," Ph.D. dissertation, University of Leeds, 2014.

[19] M. Grootendorst, "Keybert: Minimal keyword extraction with bert." 2020. [Online]. Available: https://doi.org/10.5281/zenodo.4461265

[20] M. Bani-Almarjeh and M.-B. Kurdy, "Arabic abstractive text summarization using rnn-based and transformer-based architectures," *Information Processing & Management*, vol. 60, no. 2, p. 103227, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306457322003284

[21] E. M. B. Nagoudi, A. Elmadany, and M. Abdul-Mageed, "AraT5: Text-to-text transformers for Arabic language generation," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 628–647. [Online]. Available: https://aclanthology.org/2022.acl-long.47

[22] L. Lupo, M. Dinarelli, and L. Besacier, "Focused concatenation for context-aware neural machine translation," in *Proceedings of the Seventh Conference on Machine Translation (WMT)*, P. Koehn, L. Barrault, O. Bojar, F. Bougares, R. Chatterjee, M. R. Costa-jussà, C. Federmann, M. Fishel, A. Fraser, M. Freitag, Y. Graham, R. Grundkiewicz, P. Guzman, B. Haddow, M. Huck, A. Jimeno Yepes, T. Kocmi, A. Martins, M. Morishita, C. Monz, M. Nagata, T. Nakazawa, M. Negri, A. Névéol, M. Neves, M. Popel, M. Turchi, and M. Zampieri, Eds. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, Dec. 2022, pp. 830–842. [Online]. Available: https://aclanthology.org/2022.wmt-1.77

[23] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[24] M. Post, "A call for clarity in reporting BLEU scores," in *Proceedings of the Third Conference on Machine Translation: Research Papers*. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 186–191. [Online]. Available: https://aclanthology.org/W18-6319

[25] R. REN, L.-l. ZHANG, Z.-l. GUO, and J.-l. ZENG, "The application of deverbalisation in interpreting notes," *Sino-US English Teaching*, vol. 15, no. 4, pp. 209–212, 2018.

[26] X. Tan, L.-Y. Zhang, and G.-D. Zhou, "Document-level neural machine translation with hierarchical modeling of global context," *Journal of Computer Science and Technology*, vol. 37, no. 2, pp. 295–308, 2022.