AraSim: Optimizing Arabic Dialect Translation in Children and Literature with LLMs and Similarity Scores

Alaa Bouomar
University of Leeds
Woodhouse, Leeds LS2 9JT,
UK
od21ahb@leeds.ac.uk

Noorhan Abbas
University of Leeds
Woodhouse, Leeds LS2 9JT,
UK
n.h.abbas@leeds.ac.uk

Abstract

The aim of this study is to apply advanced neural machine translation (NMT) models to translate children's stories from Modern Standard Arabic (MSA) to the Egyptian (Cairo) dialect, addressing the significant linguistic gap faced by young Arabic speakers. It utilizes state-of-the-art transformer-based models, including Claude [1] for initial translation and a fine-tuned AraT5 for back-translation and evaluation, with the aim of improving the accessibility and enjoyment of children's literature for young Egyptian readers. We assess translation quality using semantic similarity and BLEU scores. The basic AraT5 model achieved an average semantic similarity score of 0.94. Through fine-tuning on an Egyptian (Cairo) dialect-specific dataset, we enhanced these metrics, with the fine-tuned model achieving an average semantic similarity score of 0.97, representing an improvement of 3 percent. Our research has produced a high-quality parallel corpus of 130 stories, a valuable resource for future research in Arabic dialect translation. This work contributes to bridging the linguistic gap for the Arabic language between MSA and regional dialects, offering critical insights and practical solutions to enhance the educational and cultural experiences for the young Arabic speakers. The findings demonstrate significant improvements in translation accuracy and quality.

Keywords: Neural Machine Dialect Translation, Modern Standard Arabic, Egyptian Arabic Dialect, Children's Literature, Transformer Models, AraT5, Claude, Arabic NLP

1 Introduction

Modern Standard Arabic (MSA) serves as the formal and standardized form of Arabic used in written texts and formal communication across the Arab world (Obeid et al., 2020). Despite its widespread use, MSA differs markedly from the various regional dialects spoken in daily life, such as Egyptian, Gulf, Levantine, and Maghrebi Arabic (Qwaider et al., 2018). These dialects, being more prevalent in everyday conversations, informal settings, and local media, present a significant linguistic challenge for young Arabic speakers who are primarily exposed to their native dialects at home and in their communities. For children, this linguistic discrepancy could be a challenge. MSA's complex grammatical structures, extensive vocabulary, and formal style can be tough for young learners accustomed to the dialects spoken at home. This gap can affect their comprehension and enjoyment of texts written in MSA, thereby impacting their educational and cultural experiences (Al-Sulaiti et al., 2016). Addressing it is crucial for enhancing the accessibility and engagement of children's literature among young Arabic readers. This paper explores the application of advanced neural machine translation (NMT) models to bridge the linguistic gap between MSA and the Egyptian (Cairo) dialect, focusing specifically on children's literature. By employing state-of-the-art transformer-based models, including Claude for initial translation and a fine-tuned AraT5 for back-translation and evaluation, this research aims to produce high-quality translations that make children's stories more relatable and enjoyable for young Egyptian readers. The evaluation of translation quality through semantic similarity and BLEU scores further contributes to refining the models and enhancing their performance. Through this work, a high-quality parallel corpus of 130 children's stories will be developed, providing a valuable resource for future research in Arabic dialect translation. By making children's literature more accessible in the Egyptian dialect, this research not only bridge a critical linguistic divide but also enriches the educational and cultural experiences of young readers. This study offers significant insights in the field of Ara-

[1] https://claude.ai

Proceedings of The 4th Workshop on Arabic Corpus Linguistics (WACL-4), pages 93–102
January 20th, 2025.

bic dialect translation and sets the stage for further exploration and innovation in this area.

2 Background Research

Translating between Modern Standard Arabic (MSA) and regional dialects poses significant challenges due to linguistic variations and limited annotated corpora (Bouamor et al., 2018; Al-Sulaiti et al., 2016). Various approaches have been employed, including rule-based methods, statistical machine translation (SMT), and neural machine translation (NMT) (Obeid et al., 2020; Qwaider et al., 2018). While rule-based and SMT methods have provided foundational insights, they have limitations in capturing dialectal nuances. NMT, particularly transformer models like AraT5 (Obeid et al., 2020), has emerged as a promising approach due to its ability to handle long-range dependencies and complex linguistic structures.

2.1 LLMs for Arabic Dialects

Large Language Models (LLMs) like GPT-4 and Bard show varying proficiency with Arabic dialects. Kadaoui et al. (2023) found that while LLMs often outperform existing commercial systems for dialects with limited datasets, they still lag behind in MSA translation. Alyafeai et al. (2023) evaluated GPT-3.5 and GPT-4 on various Arabic NLP tasks, revealing improvements in performance but highlighting challenges in consistent evaluation across dialects. Al-Thubaity et al. (2023) and Mullappilly et al. (2023) emphasized the need for specialized training and dialect-specific corpora to enhance LLMs' proficiency in handling diverse Arabic dialects. Qwaider et al. (2018) highlighted the importance of creating dialect-specific corpora for improving LLMs in Arabic dialect translation and identification tasks.

2.2 Additional Dialectal Datasets

Recent studies have introduced several valuable dialectal datasets to improve Arabic dialect translation models (Abdelali et al., 2024). These include the Arabic Dialectal Tweets Corpus (Qwaider et al., 2018), MADAR Parallel Corpus (Bouamor et al., 2018), CALCS Dataset (Malartic et al., 2023) focusing on conversational Arabic, and the ArzEn Corpus (Waheed et al., 2023) for Arabic-English code-switching. These resources provide a range of real-world language usage examples, covering various dialects and linguistic phenomena, which can significantly contribute to training and evaluating translation models.

2.3 Evaluation Challenges

Evaluating LLMs for Arabic dialects faces several challenges. The significant variation among Arabic dialects complicates the development of standardized benchmarks (Alyafeai et al., 2023). Resource limitations, particularly the scarcity of high-quality annotated data for many dialects, constrain the effectiveness of LLMs. Al-Thubaity et al. (2023) found that while GPT-4 excelled in classification tasks, it struggled with generating high-quality dialectal text. Mullappilly et al. (2023) highlighted that LLMs often lack the cultural and contextual understanding necessary for accurate interpretation of dialectal Arabic, requiring significant fine-tuning on domain-specific datasets to achieve satisfactory performance in tasks like sentiment analysis and text generation.

3 Design and Methodology

This study aims to translate children's stories from Modern Standard Arabic (MSA) to the Egyptian dialect, leveraging advanced neural machine translation models. The primary objectives are to evaluate the performance of these models, improve their translation quality through fine-tuning, and create a high-quality parallel corpus for future research.

3.1 Methodological Framework

Our approach leverages two transformer-based models: AraT5 and Claude. Claude, developed by Anthropic, performs the initial MSA to Egyptian dialect translations, capitalizing on its contextual understanding. AraT5, specifically designed for Arabic, is fine-tuned for the Egyptian dialect and used for back-translation to MSA. This dual-model approach combines Claude's robust language generation with AraT5's specialized Arabic processing capabilities, aiming to produce high-quality translations tailored to the Egyptian dialect.

3.2 Data Sources

Arabic Children's Corpus: The Arabic Children's Corpus, compiled by (Al-Sulaiti et al., 2016), was inspired by the Oxford Children's Corpus. This corpus consists of 2,950 documents and nearly 2 million words, collected manually from the web . It includes a variety of genres specifically targeted at

children, featuring classic tales from "The Arabian Nights" and stories about popular fictional characters such as Goha. The corpus is of high quality and aims to facilitate studies in text classification, language use, and ideology in children's texts (Al-Sulaiti et al., 2016)

MADAR Corpus: The MADAR corpus is a collection of different parallel sentences covering the dialects of 25 cities or counties from the Arab World, in addition to English, French, and MSA. It was created by translating selected sentences from the Basic Traveling Expression Corpus (BTEC) (Takezawa et al., 2007) to the different dialects. The exact details on the translation process and source and target languages are described in (Bouamor et al., 2018).

3.3 Data Pre-processing Overview

For this study, we selected 130 stories of varying lengths from the Arabic Children's Corpus, chosen based on their moral and educational value. The preprocessing phase began with converting the collected Word documents into plain text files using automated scripts to ensure consistency across all files. Next, we employed natural language processing (NLP) tools to segment each story into individual sentences, storing them in a line-by-line format. We then used automated spell-checking tools to identify and correct spelling errors, followed by a manual review to ensure accuracy, especially for words with multiple correct forms depending on the context. As a final quality assurance measure, a subset of the cleaned and formatted data was manually reviewed by expert translators who are also native Egyptian speakers to verify the accuracy and quality of the text.

3.3.1 MADAR Corpus Pre-processing:

For fine-tuning and training purposes, we utilized the MADAR (Cairo) dataset. Initially, only the dialect corpus was available, with the source MSA later found on Hugging Face. We merged these Excel files using the V-lookup function. However, upon closer inspection by an Arabic and Egyptian native speaker, we identified several inconsistencies and inaccuracies in the Cairo dialect translations. These issues ranged from minor dialectal nuances to more significant semantic discrepancies. In some cases, the dialect translation did not accurately capture the meaning of the MSA sentence:

- **MSA:** أريد ساعة مستعملة.
- **Inaccurate CAI:** عايز مع ساعة يد تانية.
- **Corrected CAI:** عايز ساعة مستعملة.

In other instances, the dialect translation missed key elements of the original sentence:

- **MSA:** أريد إجازة لمدة أسبوع واحد من فضلك.
- **Inaccurate CAI:** عايز كورس مايزيدش عن اسبوع، لو سمحت.
- **Corrected CAI:** عايز إجازة لمدة أسبوع، من فضلك.

To address these issues, we split the training corpus into two sheets, each containing approximately 4500 lines. These sheets were then reviewed by two qualified Egyptian native speakers who provided suggested translations. We used Claude to compare the original CAI dialect translations with the reviewers' translations, noting improved accuracy in the translation from MSA to Egyptian Arabic.
After addressing these inconsistencies through manual review and correction, we retrained our transformer model using the updated training corpus. This process led to a significant reduction in both training and validation loss, underscoring the critical importance of data quality in machine translation tasks. The training loss decreased from 0.0892 in the first epoch before corpus improvement to 0.0484 after improvement. Similarly, the validation loss decreased from 0.05547 to 0.0360 in the first epoch. This trend of improved performance continued throughout the training process, demonstrating the value of our rigorous data preparation and correction efforts.

3.4 Transformer Models

ARAT5 Model: Building upon the work of (Nagoudi et al., 2022), we selected the AraT5 base model (PRAli22/arat5-base-arabic-dialects-translation[2]) as our starting point. This model was selected due to its demonstrated effectiveness in handling both Modern Standard Arabic (MSA) and various Arabic dialects. Our goal was to fine-tune this model specifically for translation between MSA and the Cairo dialect.

[2] https://huggingface.co/PRAli22/arat5-arabic-dialects-translation

AraT5 employs a sequence-to-sequence (seq2seq) framework with an encoder-decoder structure. Both the encoder and decoder use self-attention mechanisms and multi-head attention layers to capture dependencies and contextual information across the entire input sequence (Vaswani et al., 2017).

Claude Model: Claude is an advanced large language model (LLM) developed by Anthropic, designed to understand and generate human-like text. It is based on the transformer architecture, using self-attention to efficiently process and understand text context and dependencies (Vaswani et al., 2017). The model is pre-trained on a vast corpus of text data from diverse sources, allowing it to learn a wide range of language patterns, facts, and nuances. This extensive training helps Claude generate coherent and contextually relevant text across various topics (Brown et al., 2020).

3.5 Rationale for Utilizing Claude for Initial Translation

The choice of Claude for the initial translation from Modern Standard Arabic (MSA) to Egyptian Cairo dialect was informed by several factors. Recent research has highlighted limitations in ChatGPT's handling of Arabic dialects. Kadaoui et al. (2023) found that while ChatGPT performed well with Classical Arabic and MSA, its accuracy dropped significantly when dealing with dialectal Arabic. Claude's architecture is optimized for better contextual understanding and linguistic nuances, which are critical for accurately translating dialects (Mullappilly et al., 2023). Moreover, Claude's training incorporated a more balanced dataset including substantial representations of various Arabic dialects, potentially making it more robust for dialect-specific tasks (Waheed et al., 2023).

We conducted experiments comparing Claude and ChatGPT, with results reviewed by a native Egyptian speaker. Claude consistently outperformed ChatGPT in dialect translation tasks. Additionally, when presented with long paragraphs, ChatGPT tended to lose coherence in line-by-line translation, while Claude maintained consistency throughout. These factors, combined with Claude's demonstrated capabilities in handling complex linguistic tasks, made it the preferred choice for our initial MSA to Egyptian dialect translations.

3.6 Rationale for Dialect Translation

Young readers often find MSA challenging due to its complex grammar and formal tone. Translating children's literature to their native dialect fosters greater engagement and comprehension, creating a gateway to literature, especially in the underdeveloped countries. While MSA proficiency remains critical, introducing stories in a familiar dialect can nurture a love for reading and gradually bridge the gap to MSA.

3.7 Evaluation Metrics

To assess the quality and accuracy of our translations, we employed two primary metrics: Semantic Similarity and BLEU (Bilingual Evaluation Understudy) Score. These metrics offer complementary insights into the performance of our translation model.

Semantic Similarity: We utilized cosine similarity to quantify the semantic congruence between the original MSA sentences and their translations into Egyptian Arabic. This method allows us to measure how effectively the meaning of the original text is preserved in the translation, regardless of specific lexical choices.**:**

1. **Embedding Generation**: Both the original MSA sentences and their Egyptian Arabic translations were encoded into high-dimensional vectors using the selected model. This process transforms the textual data into a format that can be mathematically compared.
2. **Cosine Similarity Calculation**: We computed the cosine similarity between the vector representations of the original and translated sentences. This yielded similarity scores ranging from -1 to 1, where:
    - A score of 1 indicates identical semantic meaning
    - A score of 0 suggests no semantic similarity
    - A score of -1 implies opposite meanings (rarely observed in translation contexts)

BLEU Score: BLEU (Bilingual Evaluation Understudy) is a widely used metric in machine translation that compares a candidate translation to one or more reference translations. It primarily measures the precision of n-grams (typically up to 4-grams) in the candidate translation with respect to the reference(s). We use smoothing (specifically,

method4 from SmoothingFunction) to handle cases where there are no matching n-grams, which is particularly important when evaluating short sentences.

We implemented BLEU score calculation using the Natural Language Toolkit (NLTK) library in Python, which provides a robust implementation of this metric.

1. **Data Preparation**: We loaded our translated sentences along with their original MSA versions from an Excel file. The MSA sentences served as reference translations, while the model-generated Egyptian Arabic translations were the candidates for evaluation.
2. **Tokenization**: Both reference (MSA) and candidate (Egyptian Arabic) sentences were tokenized into words using simple space-based splitting. This approach assumes that words in both MSA and Egyptian Arabic are space-separated, which is generally true for written Arabic.
3. **Score Computation**: For each sentence pair, we computed the BLEU score using the tokenized reference and candidate translations. The scores were calculated individually for each sentence, allowing for a granular analysis of translation quality across our dataset.
4. **Significance in Dialect Translation:** The BLEU score provides several key insights in the context of MSA to Egyptian Arabic translation including N-gram Precision that measures how many of the n-grams (typically up to 4-grams) in the candidate translation appear in the reference translation. This is particularly useful for assessing how well the model preserves common phrases and linguistic structures from MSA to Egyptian Arabic and Brevity Penalty for translations that are too short, which helps in identifying cases where the dialect translation might be oversimplified or incomplete compared to the MSA original.

3.8 Motivation for Fine-Tuning AraT5

Initial testing of AraT5 revealed areas for improvement, prompting fine-tuning to enhance accuracy. The process involved dataset preparation, train-test split, data preprocessing, and hyperparameter optimization. The dataset was split into training (80%) and testing (20%) sets, with careful attention to tokenization, sequence length, and key hyperparameters such as learning rate and batch size.

3.9 Manual Evaluation Criteria

To ensure translation quality, we conducted manual evaluations for translations with semantic similarity scores below 96%. The evaluation focused on the following criteria:

- **Semantic Accuracy:** Faithfulness of the translation to the original meaning.
- **Dialectal Fidelity:** Appropriateness of dialect-specific expressions.
- **Cultural Relevance:** Maintenance of cultural context and tone.

Two native Egyptian speakers with expertise in linguistics independently reviewed the translations. Discrepancies were resolved through discussion.

3.10 Dealing with Different Story Sizes

To address challenges with longer stories and maintain consistency, a line-by-line translation approach was implemented for all stories regardless of length. This method helped mitigate issues of repetition and irrelevant content generation in longer texts, ensured consistent methodology across the corpus, and facilitated the creation of a structured parallel corpus for future research.

4 Experimental Results and Discussion

This section presents the findings from the translation experiments and fine-tuning processes using the Claude and AraT5 models. The results are evaluated using semantic similarity and BLEU scores to assess the quality and accuracy of translations from Modern Standard Arabic (MSA) to the Egyptian (Cairo) dialect.

4.1 Dialect-Translation-(Quality Improvement through Fine-Tuning)

The initial translations generated by AraT5 demonstrated a robust contextual understanding and high-quality output, which set a strong foundation for further refinement. However, several issues, including semantic substitution errors and inconsistencies in handling dialect-specific expressions, highlighted the need for fine-tuning the AraT5 model:

Semantic Substitution Errors: In some instances, AraT5 replaced words with semantically unrelated terms, significantly altering the meaning of sentences. A striking example of this is:

- **Original MSA** كانَ ضفدعٌ منهمكاً في صيد البعوض بلسانهِ
- **English:** A frog was busy catching mosquitoes with its tongue
- **Claude's Egyptian:** كان في ضفدع مشغول بصيد الناموس بلسانه
- **Basic AraT5:** كان كلب مشغولا في صيد الكلاب بلسانه
- **English back-translation:** A dog was busy catching dogs with its tongue
- **Fine-Tuned AraT5:** كان الضفدع مشغولا في صيد البعوض بلسانه
- **English back-translation:** The frog was busy catching mosquitoes with its tongue

Here, the basic AraT5 model replaced "ضفدع" (frog) with "كلب" (dog), and "البعوض" (mosquitoes) with "الكلاب" (dogs). Despite the high similarity score of 97.2%, this error fundamentally changes the meaning of the sentence. However, the fine-tuned AraT5 model correctly translates the sentence, preserving the original meaning.

Dialect-Specific Challenges: AraT5 sometimes struggles with dialect-specific expressions or vocabulary:

- **Original MSA:** خرج الديك وأخذ يبحث على ابنه الصغير ، فوجد الصياد يركض وبيده بيضة ، أخذ الديك يركض وراءه
- **English:** The rooster went out and started searching for his little son, then he found the hunter running with an egg in his hand, so the rooster started running after him
- **Claude's Egyptian:** خرج الديك وبدأ يدور على ابنه الصغير، فلقى الصياد بيجري وفي ايده بيضة، خد الديك يجري وراه
- **Basic AraT5:** خرج الديك و يبحث عن إبنه الصغير ، وجد الكلب يركض و به بيضة ، أخذ الحمام يركض وراءه
- **English back-translation:** The rooster went out and searched for his little son, found the dog running with an egg, and the pigeon started running after him
- **Fine-Tuned AraT5:** خرج الديك و يبحث عن إبنه الصغير ، وجد الصياد يركض بداخل البيضة ، أخذ الديك يركض وراءه
- **English back-translation:** The rooster went out and searched for his little son, found the hunter running inside the egg, and the rooster started running after him

Here, the basic AraT5 model mistranslates "الصياد" (hunter) as "الكلب" (dog) and "الديك" (rooster) as "الحمام" (pigeon) in the last part. The fine-tuned AraT5 model, while still not perfect, maintains the correct translation of "الديك" (rooster) throughout the sentence.

These examples demonstrate the significant improvement in translation quality achieved by fine-tuning the AraT5 model. The fine-tuned model is better able to handle semantic preservation, maintain consistency across translations, and navigate dialect-specific challenges. While not flawless, the fine-tuned AraT5 model represents a substantial step forward in our ability to accurately convert Egyptian Arabic dialect to MSA.

4.2 Limitations and Considerations:

It is important to note that while these metrics provide valuable insights, they each have limitations. Semantic similarity might not capture nuances in dialect-specific expressions, while BLEU scores can sometimes undervalue semantically correct translations that use different wording than the reference. To address these limitations, we also conducted manual evaluations by Arabic language experts to ensure the quality and appropriateness of the translations, especially for dialect-specific expressions and cultural nuances that automated metrics might miss.

4.3 Transformer Performance Comparison

To evaluate the effectiveness of the fine-tuning process, we compared the performance of the Basic AraT5 model with the Fine-Tuned AraT5 model. The performance was measured based on the similarity score between the translated sentences and the original MSA sentences. The similarity score was calculated using the Sentence Transformer model, which computes cosine similarity scores. The key observations were as follows:

Basic AraT5 Model: The Basic AraT5 model shows a relatively even distribution of similarity rates across the spectrum, with a slight peak around the 96% similarity score. However, it struggles to achieve higher similarity scores consistently. The Basic AraT5 model has more occurrences at lower similarity scores (88% to 89%) compared to the Fine-Tuned AraT5 model, indicating the effectiveness of the fine-tuning process.

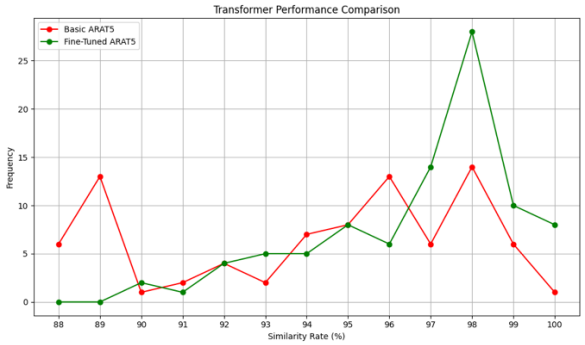

Fig. 1. Similarity Count Comparison between Basic AraT5 and Fine-Tuned AraT5

Fine-Tuned AraT5 Model: The Fine-Tuned AraT5 model demonstrates a significant improvement, with a higher frequency of translations achieving similarity scores between 96% and 100%. The peak at 98% indicates that the fine-tuning process has effectively enhanced the model's performance, resulting in translations that are closer to the original MSA sentences.

4.4 Fine-Tuned AraT5 Model

The AraT5 model was trained in two distinct phases, once before validating the training and testing datasets, and once after this validation. The primary aim was to observe the impact of dataset validation on the model's performance, particularly in terms of training loss and validation loss.

BLEU Score Analysis: The fine-tuned AraT5 model achieved a BLEU score improvement from 0.082 to 0.087. Although modest, this increase reflects the model's enhanced ability to produce linguistically accurate translations. BLEU's limitations in capturing dialectal nuances underscore the importance of complementing it with semantic similarity and manual evaluations.

4.5 Challenges with Long Stories and Line-by-Line Approach

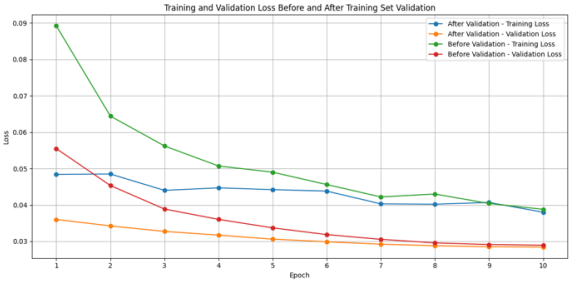

Fig. 2. Comparison of both training and validation losses after validating the training and testing datasets

Our research revealed challenges in translating longer children's stories from Modern Standard Arabic (MSA) to Egyptian dialect using neural machine translation models. Key issues included repetition of sentences in longer stories and generation of irrelevant content in very long stories. To address these challenges and ensure consistency, we implemented a line-by-line dialect-translation approach for all stories, regardless of length. This method maintained methodological consistency across the corpus, facilitated the creation of a parallel corpus, and preserved context in each line. It also ensured consistent corpus quality across all story lengths, improved coherence, and reduced irrelevant content. The method also better-preserved original content, enhanced control over translation quality, and proved scalable for handling stories of varying lengths.

4.6 Corpus Creation

We developed a high-quality parallel corpus using 130 children's stories from the Arabic Children's Corpus (Al-Sulaiti et al., 2016). Our translation process involved line-by-line translation using our fine-tuned model, followed by a selective review where human experts examined lines with semantic similarity below 96%. The resulting corpus serves as a training resource for machine translation models, an evaluation benchmark, a tool for linguistic analysis, and a unique resource for children's literature translation. To facilitate further research and development in Arabic dialect translation, we will make this corpus publicly

available on GitHub[3] in Excel format. The Excel file will include a sample of the original MSA text, the translated Egyptian (Cairo) dialect text, and the corresponding semantic similarity scores for each line. This comprehensive dataset will allow researchers to analyze the relationship between the original and translated text, as well as the quality of translations as measured by semantic similarity.

5 Conclusion

This study has made significant progress in addressing the complex challenge of translating children's stories from Modern Standard Arabic (MSA) to the Egyptian (Cairo) dialect using advanced neural machine translation models. The study's methodology combining initial translations with Claude and fine-tuning the AraT5 model, has yielded marked improvements in translation quality, as evidenced by higher similarity and BLEU scores. The basic AraT5 model achieved an average semantic similarity score of 0.945 and through fine-tuning on an Egyptian (Cairo) dialect-specific dataset, we were able to enhance this alignment, achieving an average semantic similarity score of 0.971. The 2.6% improvement demonstrates the model's enhanced capability to understand and replicate the meanings and linguistic characteristics specific to the Egyptian (Cairo) dialect, further bridging the gap between MSA and the dialect used in everyday communication by Egyptian children. Furthermore, the BLEU score, which measures the precision of the translated output by comparing it to reference translations, also showed notable improvement by increasing from 0.0828 to 0.0867 after fine-tuning. This enhancement highlighted the fine-tuned model's ability to produce translations that are not only more semantically accurate but also more aligned with human linguistic expression. Another achievement of this study is the creation of a comprehensive parallel corpus comprising 130 children's stories in both MSA and Egyptian dialect. This corpus stands to benefit future studies and model training efforts in the field of Arabic dialect translation. The study's evaluation framework, incorporating both semantic similarity and BLEU scores, along with manual reviews, ensured a thorough assessment of translation quality while maintaining cultural and contextual accuracy. This approach has resulted in translations that are culturally resonant and linguistically.

Additionally, this research contributes to enhancing the accessibility of children's literature for young Arabic speakers. By providing stories in a more familiar dialect, opening new avenues for educational and cultural engagement. In conclusion, this research represents a significant step forward in Arabic dialect translation. Its findings and resources contribute to further advancements in this crucial area of Arabic Language processing and cultural preservation.

6 Future Work

Expansion to Other Dialects: The methodology developed in this study can be extended to other Arabic dialects, such as Levantine and Gulf Arabic. Creating parallel corpora and fine-tuning models for these dialects would further bridge the linguistic gap and enhance accessibility across the Arab world.

Genre Diversification: Future work could involve expanding the corpus to include a wider range of genres beyond children's literature, such as educational materials, and popular fiction texts. This diversification would provide a broader application of the developed models and resources.

Data Augmentation: To address the limited corpus size, we plan to use data augmentation techniques such as back-translation and paraphrasing. Additionally, incorporating synthetic data generated by LLMs could enhance the diversity and quantity of training examples.

References

Latifa Al-Sulaiti, Noorhan Abbas, Claire Brierley, Eric Atwell, and Abdulmohsen Alghamdi. 2016. Compilation of an Arabic children's corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1808–1812, Portorož, Slovenia. European Language Resources Association (ELRA). https://aclanthology.org/L16-1285

Abdulrahman Al-Thubaity, Abdullah Al-Khateeb, Bassam Al-Salhi, and Mohammed Al-Ghamdi. 2023. Evaluating ChatGPT and Bard AI on Arabic sentiment analysis. *Computing Research Repository*, arXiv:2305.14745.

[3] https://github.com/alaabouomar/Optimizing-Arabic-Dialect-Translation-for-Children-s-Literature-Using-Neural-Models.git

https://www.semanticscholar.org/paper/d4c0ee9f7ea7451216845c851d069dff95545faa

Zaid Alyafeai, Anwar Al-Omari, Iman Al-Kindi, Samah Al-Riyami, and Alaeddin Al-Maqaleh. 2023. Taqyim: evaluating Arabic NLP tasks using ChatGPT models. *Computing Research Repository*, arXiv:2305.14849. https://www.semanticscholar.org/paper/d14aa448b17fdc8d4ea12b43ee1a2b1254c38703

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and Kemal Oflazer. 2018. The MADAR Arabic dialect corpus and lexicon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA). https://aclanthology.org/L18-1535

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and Sandhini Agarwal. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, pages 1877-1901. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

Taoufik Kadaoui, Houda Bouamor, Fathi Badran, and Nizar Habash. 2023. TARJAMAT: evaluation of Bard and ChatGPT on machine translation of ten Arabic varieties. *Computing Research Repository*, arXiv:2305.14786. https://www.semanticscholar.org/paper/796b894c4e1a3cb46715cc0b45a39e91ee5910e6

Quentin Malartic, Hamza Alobeidli, Danilo Mazzotta, Gabriel Penedo, Giulia Campesan, Muhammad Farooq, Mansoor Alhammadi, Julien Launay, and Badr Noune. 2023. AlGhafa evaluation benchmark for Arabic language models. In *Proceedings of ArabicNLP 2023*. Association for Computational Linguistics. https://aclanthology.org/2023.arabicnlp-1.1

Ritu Mullappilly, Mohammed Al-Awlaqi, Saleh Al-Yami, and Faisal Al-Dossary. 2023. Arabic Mini-ClimateGPT: a climate change and sustainability tailored Arabic LLM. *Computing Research Repository*, arXiv:2305.14824. https://www.semanticscholar.org/paper/6da8e97de0981b867b1038e12e98608928ad4c0e

El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2022. 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)*, pages 628-647, Dublin, Ireland. Association for Computational Linguistics. https://aclanthology.org/2022.acl-long.46/

Ossama Obeid, Nasser Zalmout, Dima Taji, Salam Khalifa, Bashar Alhafni, Koichi Inoue, Fadhl Eryani, and Nizar Habash. 2020. CAMeL tools: an open source Python toolkit for Arabic natural language processing. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 7022–7032, Marseille, France. European Language Resources Association. https://www.semanticscholar.org/paper/995ec006ac98a697ea38bd4eea8c1f3170a8adb4

Chatrine Qwaider, Nizar Habash, Houda Bouamor, and Fathi Badran. 2018. Shami: a corpus of Levantine Arabic dialects. *Computing Research Repository*, arXiv:1805.05190. https://www.semanticscholar.org/paper/654af2f5747126447e5d8fce220c6a1915761143

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, pages 5998-6008. Curran Associates, Inc. https://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf

Ahmed Waheed, Muhammad Abdul-Mageed, El Moatez Billah Nagoudi, Badr Noune, Amir Hamdi, AbdelRahim Elmadany, and Massimo Poesio. 2023. GPTAraEval: a comprehensive evaluation of ChatGPT on Arabic NLP. *Computing Research Repository*, arXiv:2305.14976. https://arxiv.org/abs/2305.14976

Ahmed Abdelali, Hamdy Mubarak, Shammur Chowdhury, Maram Hasanain, Bekhzod Mousi, Sofiane Boughorbel, Said Abdaljalil, Yasir El Kheir, Dema Izham, Fahim Dalvi, Mohamad Hawasly, Nada Nazar, Youssef Elshahawy, Ahmed Ali, Nadir Durrani, Natasa Milic-Frayling, and Firoj Alam. 2024. LAraBench: benchmarking Arabic AI with large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 487-520, St. Julian's, Malta. Association for Computational Linguistics. https://aclanthology.org/2024.eacl-long.30

7 Limitations

Dialectal Focus: The current study primarily focuses on the Egyptian (Cairo) dialect, which may not fully represent the linguistic diversity within Egypt or the broader Arab world.

Corpus Size: While the study utilized 130 children's stories, a larger corpus could provide

more comprehensive insights and potentially improve model performance.

Lack of Standardized Benchmarks: The absence of standardized benchmarks for MSA to Egyptian dialect translation makes it challenging to compare results directly with other studies.

Contextual Coherence in Line-by-Line Translation: While the line-by-line translation approach ensured consistency, it occasionally fragmented the narrative flow. To quantify this limitation, we conducted manual evaluations of story-level coherence. Future work could involve incorporating context windows to address this issue.

8 Ethical Considerations

Cultural Appropriateness: Children's literature often contains cultural elements, moral lessons, and linguistic nuances that require careful handling during translation. Relying solely on automated translation and evaluation risks misrepresenting or losing these crucial cultural elements. Human reviewers can ensure that the translations are not only linguistically accurate but also culturally appropriate and engaging for the target audience.

Age Appropriateness: Ensuring that the translated content is age-appropriate is critical, particularly when dealing with children's literature. The translations must maintain the original intent, tone, and level of complexity suitable for the target age group. This includes avoiding any content that could be deemed inappropriate or too complex for young readers.