PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA

MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH

UNIVERSITY MUSTAPHA STAMBOULI OF MASCARA

Faculty of Exact Sciences

Department of Computer Science

Dissertation

Submitted in partial fulfilment of the requirements for Doctorate degree in Computer Science

Option: Artificial Intelligence

Theme

Arabic Machine Translation of Social Media Content

Presented by **Baligh BABAALI**

Jury:

President	Omar SMAIL	Professor	University of Mascara
Director	Mohammed SALEM	Professor	University of Mascara
Examiner	Mohammed REBBAH	Professor	University of Mascara
Examiner	Mohamed Elhadi RAHMANI	MCA	University of Saida

July 15, 2024

Acknowledgements

In the beginning, thanks to Allah, the most gracious and merciful who guide and protect me on my road. The PhD was a period of intense personal growth, and I am grateful to so many people for being a part of this journey.

I want to start by expressing my sincere thanks to the jury members for their invaluable feedback and support during the evaluation process. I am truly grateful to them.

Specifically, I want to offer a heartfelt appreciation to my supervisor, **Pr. Mohammed SALEM**, whose guidance, encouragement, and mentorship have been indispensable during my thesis journey. His expertise and wisdom have played a pivotal role in guiding me to accomplish my research goals and make a meaningful contribution to the field.

I would like to thank my past mentors, professors, and educators who have enriched my learning experiences, both within and beyond the academic setting. I consider myself fortunate to have encountered numerous inspiring individuals throughout my journey, whose enthusiasm and expertise have been instrumental in shaping my growth and supporting my pursuits. Then, thanks to my family where I found the help whenever I needed it. Thanks to my friends where I found great ideas that helped me to build up this work.

Thus, thank you all for this awesome life for all the values you taught me, thank you.

Dedication

I dedicate this work to my parents:

May they find here the testimony of my deep gratitude and acknowledgment

To my wife who has always believed in me and supported me in every way possible

To my daughters, and my family who give love and liveliness.

To all those who have helped me - directly or indirectly - and those who shared with me the emotional moments during the accomplishment of this work and who warmly supported and encouraged throughout my journey.

To all my friends who have always encouraged me, and to whom I wish more success.

Thanks!

Baligh BABAALI

ملخص

تعتبر الترجمة الآلية أداة حاسمة لتحطيم حواجز اللغة وتيسير التواصل والوصول إلى المعلومات عبر سياقات لغوية متنوعة. ومع ذلك، تعتمد فعاليتها بشكل كبير على توافر كمية كافية وعالية الجودة من البيانات التدريبية، وهو تحدي يواجه في كثير من الأحيان في إعدادات اللغات منخفضة الموارد. في هذه الدراسة، نستكشف الطرق لتحسين نظم الترجمة الآلية العصبية من خلال توظيف تقنيات زيادة البيانات للتغلب على التحديات التي يواجهها مثل هذه السيناريوهات.

تشمل تجربتنا استراتيجيات زيادة البيانات المختلفة، بما في ذلك الترجمة العكسية، والنص المستنسخ، والطرق المبتكرة مثل تقنية زيادة الدوران إلى اليمين، بهدف إثراء بيانات التدريب وتحسين جودة الترجمة. من خلال تقييم دقيق مقارنة بين نماذج الترجمة الآلية العصبية المحسّنة مع النموذج الأساسي، لاحظنا تحسينات كبيرة في جودة الترجمة، كما تظهره النتائج المحسّنة لمعايير بلو. يؤكد تحليلنا فعالية تقنيات الزيادة المختلفة في تعزيز نظم الترجمة الآلية العصبية، خصوصاً في سياقات اللغات ذات الموارد المنخفضة.

علاوة على ذلك، يسلط تحليلنا المقارن بين نماذج الترجمة الآلية العصبية من نوع تسلسلي-تسلسلي والنماذج المبنية على الجي.بي.تي الضوء على تعقيدات هياكلها المعمارية وسمات أدائها. من خلال تقييم أدائها في مهام الترجمة المتنوعة، وجدنا أن نموذج شات جي.بي.تي يتفوق بشكل متسق على النموذج تسلسلي-تسلسلي، حيث يظهر معايير كومت و بلو و سي.ايتش.ار.اف أعلى. بشكل لافت، أظهر نموذج شات جي.بي.تي أداءً متفوقًا في الترجمة من اللهجة العربية الجزائرية إلى العربية الفصحى الحديثة. علاوة على ذلك، أدى الانتقال من السيناريوهات صفر-مساعدة إلى قليلة-المساعدة إلى تحسين أداء الترجمة لنماذج شات جي.بي.تي عبر كل من زوجي اللغات سابقتي الذكر. تساهم هذه النتائج في فهم أعمق للتفاعل بين نماذج شات جي.بي.تي والنماذج المبنية على جي.بي.تي في مجال الترجمة الآلية، مما يمهد الطريق لتطورات مستقبلية في هذا المجال.

كلمات مفتاحية: اللغة ذات الموارد المنخفضة، زيادة البيانات، النموذج اللغوي الكبير، الترجمة الآلية العصبية.

Abstract

Machine translation serves as a crucial tool for breaking down language barriers and facilitating communication and information access across diverse linguistic contexts. However, its efficacy heavily relies on the availability of sufficient and high-quality training data, a challenge often encountered in low-resource language settings. In this study, we explore methods to enhance Neural Machine Translation (NMT) systems by employing data augmentation techniques to address the challenges posed by such scenarios.

Our experimentation involved various augmentation strategies, including Back Translation, Copied Corpus, and innovative methods like Right Rotation Augmentation, with the aim of enriching training data and improving translation quality. Through rigorous evaluation comparing augmented NMT models with the baseline, we observed significant enhancements in translation quality, as evidenced by improved BLEU scores. Our analysis underscores the effectiveness of different augmentation techniques in bolstering NMT systems, especially in low-resource language contexts.

Furthermore, our comparative analysis between Seq2Seq NMT models and GPT-based models sheds light on their architectural intricacies and performance characteristics. Evaluating their performance across diverse translation tasks, we found that the ChatGPT model consistently outperformed the Seq2Seq model, exhibiting higher COMET, BLEU, and ChrF scores. Notably, the ChatGPT model demonstrated superior performance in translating from the Algerian Arabic dialect (DZDA) to Modern Standard Arabic (MSA). Moreover, transitioning from zero-shot to few-shot scenarios led to enhanced translation performance for ChatGPT models across both language pairs. These findings contribute to a deeper understanding of the interplay between Seq2Seq and GPT-based models in machine translation, offering valuable insights for future advancements in the field.

Key words: *Neural Machine Translation, Large Language Model, Data augmentation, Low resource language.*

Contents

List of Figures

List of Tables

List of Algorithms

List of Abbreviations

BART:	Bidirectional Auto-Regressive Transformers
BERT:	Bidirectional Encoder Representations from Transformers
BLEU:	Bilingual Evaluation Understudy
BRNN:	Bidirectional Recurrent Neural Network
BT:	Back Translation
FFN:	Feed-Forward Network
GPT:	Generative Pre-trained Transformer
GRU:	Gated Recurrent Units
LLM:	Large Language Model
LSTM:	Long Short-Term Memory
MT:	Machine Translation
NLG:	Natural Language Generation
NLP:	Natural Language Processing
NLU:	Natural Language Understanding
NMT:	Neural Machine Translation
RBMT:	Rule-Based Machine Translation
RNN:	Recurrent Neural Network
Seq2Seq:	Sequence to Sequence
SMT:	Statistical Machine Translation
T5:	Text-to-Text Transfer Transformer

Chapter 1

Introduction

Machine translation (MT) serves as a vital tool for facilitating access to information, enabling cross-language information retrieval, breaking cultural and language barriers and aiding in speech interpretation. Essentially, it serves to break down language barriers that could otherwise lead to social isolation. Neglecting languages with fewer resources can have severe consequences for societal integration in today's interconnected world. Moreover, it increases the risk of digital language extinction, a phenomenon exacerbated by the digital divide [5].

The complexity of MT stems from various factors, including morphological differences among languages [6]. Additionally, categorizing languages for comparative analysis poses challenges [7]. Various methods have been employed to automate translation, initially relying on rule-based systems. However, creating such systems is laborious and expensive due to the challenges of encoding all necessary language knowledge accurately. Additionally, it requires extensive linguistic expertise and resources, which may be lacking for low-resource languages [8]. Consequently, data-driven or corpus strategies gained traction as access to parallel corpora increased. These approaches leverage curated parallel training data to develop translation models through machine learning. Among corpus approaches, Statistical Machine Translation (SMT) and Neural Machine Translation (NMT) are prominent. NMT has surpassed SMT in recent years due to several advantages. Unlike SMT, NMT allows for joint optimization of all system components to enhance translation performance. It processes complete sentences rather than just words or n-grams, leading to better handling of syntactic and semantic differences between languages. Ultimately, NMT produces more fluent translations compared to SMT [9, 10, 11, 12].

However, NMT has certain limitations, as detailed in Section 1.1, regarding low-resource languages. Therefore, we adopted the design science research methodology [13] to design, build, and evaluate an NMT system that suits low-resource languages. The design science process includes problem identification and motivation (Section 1.1); definition of the objec-

tives for a solution (Section 1.2); and communication (Section 1.3)

1.1 Problem Identification and Motivation

While Neural Machine Translation (NMT) has shown considerable advancements for high-resource languages, its performance tends to be lower for less-resourced languages [9, 14]. This disparity primarily stems from the quantity and quality of available training data, which significantly impacts NMT model performance [15]. However, the majority of the approximately seven thousand languages spoken worldwide lack sufficient training data with the requisite quantity and quality for effective NMT. One approach to address the challenge of scarce data involves optimizing the hyperparameters of NMT, which play a crucial role in the architecture design of NMT systems. Various optimization techniques have been proposed for NMT, each exhibiting differing levels of performance depending on the size of the training data [16]. Consequently, optimizing NMT hyperparameters for low-resource languages becomes imperative. Moreover, NMT encounters challenges related to its fixed vocabulary, primarily due to limitations in computing resources such as GPUs and memory in computers. This constraint poses difficulties for NMT models in handling rare and out-of-vocabulary words in text [17]. This issue is particularly pronounced in languages with complex morphologies, such as Arabic, where a single word may undergo numerous inflections, and the lexicon of the language may comprise hundreds of thousands or even millions of entries.

Hence, we aimed to address two main research questions:

- RQ1: Do data augmentation techniques enhance NMT scores under low data conditions?
- RQ2: Does a seq2seq NMT system outperform a GPT-based system in low-data scenarios?

1.2 Objectives and Scope

Our aim was to devise a specialized NMT framework tailored for languages lacking in resources, particularly focusing on the Algerian Arabic dialect (DZDA) to Modern Standard Arabic (MSA) translation task. We commenced this endeavour by crafting a comprehensive corpus for the Algerian Arabic dialect to Modern Standard Arabic translation. Subsequently, we engineered a Seq2Seq NMT model customized for this low-resourced language pair. Our

primary goal was to experiment with data augmentation techniques such as Back Translation, Copied Corpus, and the novel techniques we developed—Right Rotation Augmentation and Entity Replacement Augmentation—to enhance the quality of the Seq2Seq NMT model. Additionally, we aimed to design an NMT architecture suitable for low-resource languages and create a DZDA-MSA corpus. Our second objective was to develop an effective translation system capable of handling the unique challenges presented by this language pair. To evaluate the efficacy of our Seq2Seq NMT model, we compared its performance against a GPT-based machine translation model. This comparative analysis allowed us to gauge the strengths and weaknesses of each approach in the context of translating between DZDA and MSA. Through this endeavour, we aimed to contribute insights into the optimization of machine translation systems for low-resource languages and explore the capabilities of different NMT architectures in addressing translation challenges specific to such language pairs.

1.3 Contributions

Our research outcomes have been disseminated through various publications presented at prestigious scientific conferences. At the 7th International Symposium (**MISC 2022**), we introduced the "Survey of the Arabic Machine Translation Corpora" [18]. Additionally, at the 1st International Conference on Artificial Intelligence: Theories and Applications (**ICAITA 2022**), we presented "Arabic Machine Translation: A Panoramic Survey" [19]. Alongside these contributions, we curated an in-house corpus specifically tailored for testing our NMT models, which we have made openly accessible to the MT community for research purposes, thereby enhancing the replicability of our findings. Additionally, we developed innovative data augmentation techniques designed to improve the performance of NMT models. The experiments and results are detailed in our forthcoming paper, "Chasing the Recipe for Effective Low-Resource Neural Machine Translation". In our study, detailed in the article "Breaking Language Barriers with ChatGPT: Enhancing Low-resource Machine Translation between Algerian Arabic and MSA" [20], published in the esteemed International Journal of Information Technology **(IJIT)**, we performed comparative experiments and analyses between Seq2Seq and GPT-based models for machine translation tasks.

1.4 Outline of the Thesis

In Chapters 2, 3, and 4, we delve into a comprehensive exploration of the Arabic language, the fundamentals of Neural Machine Translation (NMT), a literature review encompassing NMT, Large Language Models, and relevant research in the field. Our focus on low-resource

languages necessitates meticulous preparation, involving the compilation of various corpora types and the implementation of a spelling corrector to refine the collected data. Chapter 5 provides an in-depth discussion on the compilation process of a monolingual corpus, tailored specifically for the development of a robust NMT system targeting the low-resourced language pair of Dialectal and Modern Standard Arabic. In Chapter 6, we delve into the exploration of data augmentation techniques aimed at enhancing the performance of the Seq2Seq NMT model. The subsequent chapter, Chapter 7, revolves around the development and comparison between the Seq2Seq NMT model and the GPT-based model, offering insights into their respective strengths and limitations. Finally, Chapter 8 encapsulates our concluding remarks and outlines potential avenues for future research endeavors in this domain.

PART I:
BACKGROUND AND RELATED WORK

Chapter 2

Arabic Language and Machine Translation

2.1 Overview

Translation has become difficult due to the intricate variations across languages [6]. For instance, certain words may have different meanings based on the context, or other words may not have equivalent translations in other languages. Additionally, translating idiomatic expressions calls for a thorough understanding of both the source and target languages. Further, structural variations like word order disparities between languages complicate translation.

Another difficulty is that a good translation needs to be faithful and fluent. A faithful translation accurately conveys the sense of the original text, whereas a fluent translation is easy to understand and sounds natural. A literal, faithful translation could result in an unpleasant and unnatural translation in the target language. For example, fluency rather than faithfulness is more critical when translating literary works. We might have to alter some of the meaning to keep the text flowing smoothly. Readers should feel as if it was written in their native language. The faithfulness of the translation, however, is prioritized when translating a technical manual or a legal document. Even if the translation is not fluent, it must be faithful and convey the same meaning. To produce accurate translations that balance faithfulness and fluency, human translators primarily rely on their experience, knowledge, and reasoning abilities. Due to these issues, various human translators will translate the same text in different ways.

Despite the complex linguistic distinctions, recent decades have seen significant improvements in machine translation. It is even applied in practical, real-world applications. For instance, we employ machine translation for cross-language information retrieval. It allows people to interact and obtain information in other languages. Machine translation is also

used to assist human translators. By creating a draft translation that human translators will edit, it expedites a time-consuming translation task [21]. In addition, we can employ machine translation for translations that are speech- and image-centric. Speech-centric translation involves translating a text from a speech recognition system into another language before the text is fully formed. As a result, it mimics a live human interpretation. Image-centric machine translation uses an optical character recognition system to translate the text included in images, such as billboard advertisements or street signs.

There are various methods for automating the challenging task of translation. Data-driven methods later supplanted the initial rule-based approaches. The most popular data-driven methods are Statistical Machine Translation (SMT) and Neural Machine Translation (NMT). Nevertheless, due to its remarkable successes, NMT has become state-of-the-art.

This chapter provides a comprehensive overview of the intersection between Arabic language characteristics and machine translation methodologies. It begins by delving into the intricate nuances of the Arabic language in the section 2.2, highlighting its unique linguistic features and complexities. Subsequently, the discussion extends to the role of social media in shaping Arabic language usage and its implications for natural language processing tasks in the sections 2.3 and 2.4. Within the realm of machine translation, the section 2.5 explores various linguistic and corpus-based approaches employed to facilitate translation from Arabic to other languages and vice versa. Moreover, the section 2.6 delves into the evaluation methodologies utilized to assess the effectiveness of machine translation systems, encompassing both human and automatic evaluation techniques. Throughout the section 2.7, relevant related work is examined, providing insights into previous research endeavours and advancements in the field.

2.2 Arabic Language

Arabic is considered as one of the six official languages of the United Nation. It is the official language in 22 countries and spoken by more than 350 million people in 24 countries around the world [22]. The Arabic language is morphologically rich and complex, however, it is considered a low-resource language due to the lack of enough parallel dataset.

Arabic is well known for its complex morphology. It has different possibilities of word order that express the same sentence. According to word orders, Arabic sentences can be classified into 4 types: SVO[1], VSO, VOS and SOV [23].

Translating the Arabic language into other languages engenders multiple linguistic problems, as no two languages can match, either in the meaning given to the conforming symbols

[1]SVO: Subject-Verb-Object

or in the ways in which such symbols are arranged in sentences. Lexical, syntactic and semantic problems arise when translating the meaning of Arabic words into English and vice versa. Machine translation into morphologically rich languages (MRL) poses many challenges, from handling a complex and rich vocabulary that can reach hundreds of thousands or even millions, to designing adequate MT metrics that take morphology into consideration.

Fehri[24], Chalabi[25] and Daimi[26] enumerated major issues involving Arabic in the following points:

- Arabic is written from right to left.
- There are no capital letters in Arabic.
- Gender is used for all nouns (there are no neutral).
- Some letters have different shapes depending on their location within a word. e.g. The shape of letter (ع) in the start of a word is (ع) like in علب, in the middle (ـعـ) like in لعب and at the end (ع or ـع) like in بلع or ودع.
- Arabic words can be partially, fully or not vocalized. Unvocalized words may generate ambiguities for MT. See Table 2.1.
- Arabic has a relatively free order of words. See Table 2.2.
- Some Arabic vocalized words may have multiple senses (polysemy) depending on its context. See Table 2.3.
- The subject can be omitted. e.g. يشرب الماء (He drinks the water)
- Some words hold the meaning of a whole sentence. e.g. فَأَسقَينَاكُمُوه (and We gave it to you to drink).
- Copula verbs "to be" and "to have" do not exist in Arabic.
- The three letters root system can often engender ambiguous words.
- Feminine nouns are often derived from masculine nouns, e.g. مهندس (Engineer male)

Word	Transliteration	English meaning
عَمَرَ	Eamara	build / live / was populated
عَمَّرَ	Eam˜ ara	Live a long time
عُمِرَ	Eomira	became populous
عُمِّرَ	Eom˜ ira	Given a long life
عُمَرُ	Eumaru	Omar (noun) / plural of Umrah
عَمْرُ	Eamoru	Age of
عُمْرُ	Eumoru	Age of
عُمُرُ	Eumuru	Age of
عَمْرٌ	EamorN	Age
عُمْرٌ	EumorN	Age
عُمُرٌ	EumurN	Age
عَمَرٌ	EamarN	Head cover for women
عَمَرَّ	Eamar˜ a	Strong tough man / Longest of everything / Shrewd malicious guy / Ferocious man

Table 2.1: Possible meanings of the unvocalized Arabic word "عمر" (Emr)

Order	Sentence	English translation
VSO	شربَ عمرُ الماءَ	Drank Omar the water
OVS	اَلماءَ شربَ عمرُ	The water drank Omar
SVO	عمرُ شربَ الماءَ	Omar Drank the water
VOS	شربَ الماءَ عمرُ	Drank the water Omar

Table 2.2: Arabic free word order

مهندسة (Engineer female). In some cases they are totally different, e.g. ولد (Boy) بنت (Girl).

- In English the number system moves from singular to plural form directly, however

Word	Transliteration	English meanings
سَائِل	sA}il	liquid, beggar, questioner
عَين	Eayn	eye, water source, gold

Table 2.3: Examples of vocalized Arabic words polysemy

Arabic language includes dual form by suffixing morpheme (ان) or (ين) to the singular form. e.g. مهندس (Engineer male) مهندسَان or مهندسَين (Two Engineers male).

- The plural form of Arabic masculine nouns is the result of suffixing morpheme (ون) or (ين) to the singular form. e.g. مهندس (Engineer male) مهندسُون or مهندسِين (Engineers male).
- The plural form of Arabic feminine nouns is the result of suffixing morpheme (ات) to the singular nouns. e.g. مهندسة (Engineer female) مهندسَات (Engineers female).
- Some words have no fixed rule for their plural form. e.g. طبيب (Doctor) أَطِبَّاء (Doctors).

In addition to the preceding specific challenges to MT, common standard issues also are present in Arabic language, such as:

- Multi word expressions (MWE) where the meaning of words collocation varies between partially to completely not derivable from its single constituents [27]. e.g. قَاعدة عسكرية, (military base) is a phrase that is highly compositional. مدينة الملَاهي , (amusement park), lit. "city of amusements" is an expression that shows a degree of idiomaticity. In extreme cases the meaning of the expression as a whole is utterly unrelated to the component words, such as, فرس النبي, (grasshopper), lit. "Horse of the Prophet".
- Idioms and idiomatic expressions are frequently used by Arabic speakers. They are a special challenge for MT systems, because their translation mainly does not outcome

literally, but logically [28]. e.g. أَظلَم من تِمسَاح , lit. "More oppressive than a crocodile", has the idiomatic meaning of (crocodile tears).

- Named entities (NE) refer to abstract entities in the real world such as people (PER) such as اَلأَمِير عبدالقَادر (Emir Abdelkader), places (LOC) such as مكة (Mecca), companies, and organizations that have an appropriate name (ORG) such as يَاهُو (Yahoo). It also refers to expressions of date, time, space and quantity (MISC) such as 8 سبتمبر (September 8th), 100دج (100DZD) or 25كغ (25Kgs).

  Due to the rich lexical variations and the absence of capitalization in Arabic, the task of named entities recognition (NER) is more difficult. Handling NE in Arabic MT is done very carefully based on :

  1. meaning translation: e.g. اَلأُمَم المُتَّحدَة (The United Nations).
  2. phoneme transliteration: e.g. جُوجل (Google).
  3. mixture of meaning translation and phoneme transliteration: كَارُولَاينَا الشمَاليَة (North Carolina), where "North" is translated and "Carolina" is transliterated.

2.3 Social Media

Social media involves online spaces where individuals can post and share information with others. They have become crucial in contemporary communication, fundamentally reshaping how individuals connect, information disseminates, and communities form. These online environments, characterized by user-generated content and interactive features [29], allow for the creation of virtual spaces fostering real-time interaction and knowledge sharing [30].

Social Networking Sites (SNS): Platforms like Facebook prioritize building and maintaining online social connections. Users can share personal updates, engage in discussions through comments and reactions, and participate in groups centered around shared interests [31]. Research suggests that SNS use can foster social capital by strengthening existing relationships and enabling the formation of new ones [32].

Media-Sharing Platforms: Platforms like Instagram and YouTube focus heavily on the creation and dissemination of multimedia content. Users can share and discover photos, videos, and live streams, with features designed to enhance content visibility through hashtags and algorithmic recommendations [33]. Studies highlight the increasing influence of visual communication and the emergence of new social norms surrounding content creation and consumption on these platforms [34].

Messaging Applications: Platforms like WhatsApp and WeChat enable asynchronous and real-time communication between individuals and groups. These applications offer features like text messaging, voice calls, and video conferencing, facilitating private and semi-private communication, often replacing traditional communication methods [35]. Research suggests that messaging applications can strengthen interpersonal bonds and serve as crucial tools for information sharing and community mobilization [36].

Professional Networking Sites: LinkedIn serves a distinct purpose within the social media landscape, focusing on professional networking and career development. Users create profiles highlighting their skills and experience, connect with colleagues and potential employers, and discover job opportunities [37]. Studies suggest that LinkedIn usage can positively impact career prospects by facilitating professional connections and knowledge exchange [38].

Social media platforms are a goldmine for data collection. However, in the case of Arabic, several issues arise due to the way language is used on these platforms. Users often tend to write in colloquial Arabic or a mix of colloquial Arabic and English characters (often referred to as Arabizi). This deviates from the standard Arabic used in formal writing. This informality poses challenges for data collection and analysis, as algorithms need to be able to understand these diverse and non-standard language patterns. Another important aspect of Arabic language on social media is the use of dialects. Arabic is a highly diverse language, with many different dialects and variations. Users may choose to write in their local dialect, which can vary significantly from Modern Standard Arabic, the standardized form of the language. With the spread of social media, new challenges arise to the Arabic MT such as:

- **Non-standard speech**: which encompasses dialectal languages or the various colloquial forms of standard language. This text frequently contains slang, MWE and unreasonable abbreviations [39] like as "idk[2]" and "brb[3]" in English or "hmd[4]" and "slm[5]" in Arabic. For best translation results, the MT system is required to identify such argot and try to map it to the target language.

[2] I don't know

[3] Be right back

[4] أَلحمد لِلّه (Thank God)

[5] سَلَام (Peace)

- **Arabizi** (sometimes known as Arabic chat alphabet, Franco-Arabic, Arabish, Araby and Mu'arrab): which is defined as writing informal Arabic dialects in Latin characters and Arabic numerals. This new style of writing doesn't follow any type of rules, which leads to big variations in writing nearly all Arabic words. The following words present some examples of the Arabizi issue that machine translation systems have to solve [40]:

  1. The Modern Standard Arabic word تحرير[6] has the following popular Arabizi equivalents: ta7rir, tahrir, t7rir, t7reer, ta7reer, tahreer, etc.
  2. The dialectal spellings of the MSA words لايلعب[7] could be مَايلعبش, مَابيلعبش, ميلعبش, ميلعبشِي, مَايلعبشِي etc, and the resultant Arabizi could be: mayel3absh, mayel3abchi, mabyelaabsh, mabyel3absh, mayel3abshi, mayel3abchi, etc.

2.4 Natural Language Processing

Natural Language Processing (NLP) is a branch of computer science dedicated to facilitating communication between computers and human languages. Its methodologies find application in social media for the analysis and comprehension of the extensive textual data produced by users (Huang, 2021). Below are several applications of NLP in social media:

1. **Machine Translation:** NLP techniques are utilized to translate text from one language to another, facilitating cross-lingual communication and information access. Machine translation involves converting text from a source language into a target language while preserving meaning, context, and grammatical structure.

2. **Text Classification:** NLP techniques involve categorizing text data into predefined categories or labels based on its content. This involves tasks such as sentiment analysis, spam detection, topic classification, and intent recognition in chatbots. Text classification is essential for organizing and analyzing large volumes of text data efficiently.

3. **Named Entity Recognition (NER):** NLP algorithms aim to identify and extract named entities from unstructured text, such as names of people, organizations, locations, dates, numerical expressions, and other specific entities. NER is crucial for var-

[6] Liberation
[7] He does not play

ious applications, including information retrieval, entity linking, and knowledge graph construction.

4. **Speech Recognition:** NLP algorithms transcribe spoken language into text, enabling voice-controlled applications, virtual assistants, and speech-to-text transcription services. Speech recognition involves converting audio signals into textual representations, which can then be further processed or analyzed.

5. **Information Extraction:** NLP methodologies focus on extracting structured information from unstructured text sources, enabling the retrieval of specific data points or facts from documents, web pages, or other text-based sources. Information extraction techniques involve identifying relevant entities, relationships, and events mentioned in the text.

6. **Text Generation:** NLP models are employed to generate human-like text based on given prompts or input. This includes tasks such as automated content creation, dialogue generation, language translation, and summarization. Text generation techniques leverage deep learning architectures such as recurrent neural networks (RNNs) and transformer models.

7. **Document Summarization:** NLP techniques condense lengthy documents or articles into shorter summaries while preserving the main points and key information. Document summarization aims to provide a concise overview of the text, facilitating easier comprehension and information retrieval.

8. **Question Answering:** NLP systems process natural language questions and retrieve relevant answers from large text corpora or knowledge bases. Question answering involves understanding the semantics and context of the question, searching for relevant information, and generating accurate responses.

9. **Language Understanding:** NLP models aim to understand the semantics, context, and intent behind natural language expressions. Language understanding involves tasks such as text comprehension, semantic analysis, and context modeling, enabling machines to interpret human communication more effectively.

10. **Dialog Systems:** NLP-powered conversational agents interact with users in natural language, providing assistance, answering queries, and engaging in dialogue across various domains. Dialog systems leverage techniques such as natural language understanding (NLU) and natural language generation (NLG) to enable seamless communication between humans and machines.

Natural Language Processing (NLP) emerges as a transformative force, enriching our lives by bridging the gap between human language and computing systems. This remarkable technology enables machines to comprehend, interpret, and generate human language, revolutionizing how we interact with technology. Through NLP, machines can analyze vast volumes of textual data from diverse sources such as social media, news articles, and customer reviews, extracting meaningful insights that shape various aspects of our lives. From facilitating accurate language translation and sentiment analysis to powering virtual assistants and chatbots, NLP permeates numerous applications, enhancing efficiency and convenience in our daily activities. Its significance lies in its ability to decipher the intricacies of human communication, empowering individuals and businesses alike with deeper understanding and actionable intelligence.

2.5 Machine Translation

Machine Translation is a procedure that uses computer pieces of software to express text from one natural language NL (SL i.e. source language) in another NL (TL i.e. target language). In any human or automated translation process, the meaning of the source sentences must be fully reproduced into the target translated sentences, which is only simple on the surface.

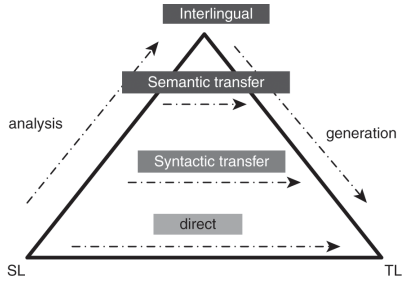

Figure 2.1: The Vauquois triangle, illustrating the foundations of machine translation.

The different approaches to MT fall into three categories: methods that depend on rules and knowledge (linguistic-based). Approaches that are empirical and data-driven (corpus-based); and finally, hybrid methods.

2.5.1 Linguistic Approaches

These MT approaches attempt to formalize all the necessary knowledge required for translation, using expert methods. The "Vauquois triangle" presented in Figure 2.1 is a generic representation of these techniques.

2.5.1.1 Direct Approach

or Direct MT (DMT) is, the simplest MT approach. It operates at the word level, i.e. the words' translation is done word by word, just, as a dictionary does, and generally without much correspondence of their meaning [41].

2.5.1.2 Rule-based MT

(RBMT) uses linguistic knowledge of source and target languages fundamentally collected from (bilingual) dictionaries and grammars encompassing the principal morphological, syntactic and/or semantic rules of each language respectively [41]. RBMT approach suffers from the impossibility of writing all the rules of all the languages, because this task requires large and important linguistic knowledge.

2.5.1.3 Interlingual MT

The term "Interlingua" refers to a language that serves as a bridge between two languages. In this method, SL is turned into an assistant/mediator language (representation) which is independent of the languages concerned by the translation. This auxiliary form is then used to specify the TL's translated verse. This approach focuses on a single representation for different languages [42].

2.5.1.4 Transfer-based MT

(TBMT) is similar to Interlingual-MT in that it generates a translation from an intermediate structure that mimics the original sentence's meaning. The source text is translated into a less language-specific intermediate representation. This form is then translated into a target language structure with a comparable structure, and the text is generated in the target language. The source and target languages' morphological, syntactic, and/or semantic information is used in the transfer process. As a result, TBMT can make use of knowledge of both the source and target languages. [43].

2.5.2 Corpus Approaches

Corpus techniques use empirical methods to ensure that all linguistic knowledge is learned empirically and automatically from corpora, which are collections of parallel datasets of source and target phrases that are translated to each other.

2.5.2.1 Example-based MT

The main idea behind (EBMT) is analogy [44]. The primary concept is to build new translations on top of current examples. Bilingual parallel corpora containing sentence pairs are used to train EBMT systems. It's used to translate similar-sounding sentences by looking for the closest source example to the source word or phrase in parallel corpora. Nagao has appropriately classified this procedure into three steps [44]:

- Fragments are matched against a database of real examples.
- Identifying the translation fragments that correlate (Alignment)
- Putting these together to create the target text

2.5.2.2 Statistical MT

(SMT) generates translation hypotheses in a target language t based on a sentence in a source language s with the highest conditional probability $P(t|s)$ [45, 46]. The translation direction will be inverted to a translation model (TM) $P(s|t)$ and a language model (LM) $P(t)$ will be included by applying the Bayes rule. The following equation (2.1) is used to optimize the likelihood of the best translation:

$$t_{best} = arg_t max(P(t|s)) = arg_t max(P(s|t) \times P(t)) \tag{2.1}$$

where $P(s|t)$ is the TM and $P(t)$ is the LM.

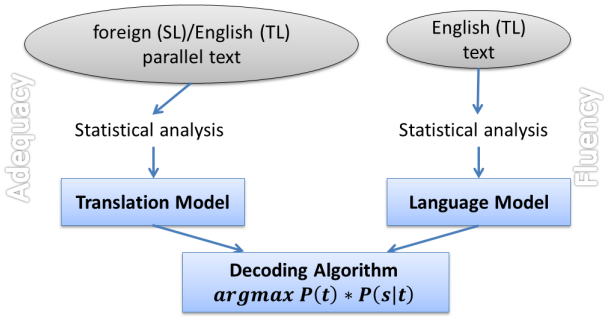

Figure 2.2: Example of the Statistical Machine Translation approach

SMT requires a language model, a translation model, and a decoding method in general. The TM, on the one hand, assures that the target hypothesis created matches the source sentence. The LM, on the other hand, ensures that the output is grammatically correct (Figure 2.2).

2.5.2.3 Neural MT

Neural Machine Translation (NMT) is fundamentally a corpus-based approach, relying heavily on large datasets of parallel texts for training. The details of this methodology, including the basics, training, and evaluation of the models, will be thoroughly explored in the next chapter.

2.6 Evaluation

In the realm of design science research, evaluation plays a pivotal role in gauging the efficacy of machine translation outputs and discerning between various translation methodologies. The assessment of quality is imperative for monitoring progress, ideally culminating in a singular metric. However, the formulation of such a metric remains an ongoing research endeavour [47]. Despite this, certain best practices have emerged, and there is generally widespread agreement on how to measure quality improvements. Both human and automatic evaluation methods are available, with human evaluation often considered more accurate, given that translations are ultimately intended for human consumption. Nonetheless, conducting human evaluations can be resource-intensive and time-consuming, making them feasible only for comparing a limited number of variant systems. Consequently, automated metrics are frequently utilized due to their ability to rapidly assess system enhancements and serve as a loss function for model training.

2.6.1 Human Evaluation

Human evaluation can be considered the most common method of judging and measuring translation quality. The linguistics and translation experts can judge the quality of a translation system output from two corners:

- **Fluency**: The level of smoothness and coherence of the translated text to target language norms, such as grammaticality, intelligibility and clarity. When annotating fluency, the evaluators (fluent only in the target language) have access to only the translation being evaluated and not the source text which is not relevant to the fluency assesment.

- **Adequacy**: Also known as accuracy. It stands for the correspondence of the target text to the source text, including the expressive means in translation, and how well the target text represents the informational content of the source text. In this case, the bilingual evaluators (in both the source and target languages) have access to the source

text and translations being evaluated and habitually, they take into consideration the context of the sentence.

The fluency and adequacy are usually measured on a 5-point scale, as presented in the following table 2.4 [48].

Adequacy	**1**	**2**	**3**	**4**	**5**
	None	Little meaning	Much meaning	Most meaning	All meaning
Fluency	**1**	**2**	**3**	**4**	**5**
	Incomprehensible	Disfluent language	Non-native language	Good language	Flawless language

Table 2.4: Numeric scale for adequacy and fluency evaluation.

Human evaluation is intrinsically subjective and expensive in time and money. To reduce the problem of subjectivity, more experts are usually invited to evaluate the same translations in the ES, and their assessments are, eventually, justified statistically.

2.6.2 Automatic Evaluation

Generally, translation evaluation methods are based on counting word- and/or sentence-based errors likely to be identified automatically. Correlation with human evaluation is the measure of evaluation for metrics. Different metrics are used in MT evaluation: BLEU, COMET, ChrF, NIST, METEOR, TER, LEPOR and many others. All of these metrics require reference translations because they confront the MT output sentences with reference translations and produce comparison scores.

2.6.2.1 BLEU

BLEU metric [49] is one of the first and most used metrics to return high correlation with human evaluation of quality. It measures the overlap of single words (unigrams) and n-grams between MT output and reference translations. BLEU operates by not only counting matching words between the translation and reference but also accounting for n-gram alignments. This approach values proper word sequencing, enhancing the chances of aligning word pairs (bigrams) or longer sequences like trigrams or 4-grams. Additionally, multiple reference translations can be employed to assess the presence of n-gram matches across variations. The BLEU score of a machine-translated output relies on the adjusted n-gram precision along with a brevity penalty. Essentially, precision measures the proportion of n-grams in the

machine translation output that align with the reference translation. BLEU scores are calculated across an entire test set, typically with one or more reference translations. However, it's uncommon in practice to utilize multiple reference translations.

$$BLEU = BP * exp \sum_{n=1}^{N} log \frac{matching\_i\_grams}{total\_i\_grams} \tag{2.2}$$

The brevity penalty (BP) in Equation 2.2, which penalizes shorter output, is expressed as:

$$BP = min(1, \frac{output\_length}{reference\_length}).$$

BLEU suffers from notable drawbacks, one of which is its stringent nature. For instance, it incorporates trigram or 4-gram precision in its calculation. However, a translated sentence might lack any trigram or 4-gram matches with the reference translation, leading to a BLEU score of zero.

2.6.2.2 ChrF

While metrics like BLEU score traditionally emphasize word-level similarity between translated text and a reference translation, this approach can pose challenges for languages with intricate morphology, where a single word in the source language may translate into multiple words in the target language, such as Arabic. To address this limitation, researchers introduced ChrF (CHaRacter-level F-score) [50] as an alternative metric.

ChrF operates at the level of character n-grams rather than word n-grams. It computes the F-score, a harmonic mean of precision and recall, for matching character sequences between the translated text and the reference translation. This focus on character-level matches enables ChrF to capture semantic equivalence even when word-level order or morphology differs.

ChrF presents several advantages over word-level metrics.

- Firstly, it proves particularly effective for assessing translations involving languages with complex morphology.
- Secondly, it is less susceptible to issues stemming from variations in word order within the translated text.
- Additionally, ChrF can be combined with word-level n-gram metrics to furnish a more comprehensive evaluation, leveraging the strengths of both approaches.

Nevertheless, ChrF also exhibits limitations. As it concentrates solely on character sequences, it may not consistently account for overall fluency or grammatical accuracy in the

translated text. Furthermore, the selection of the n-gram size can influence the score, posing challenges in determining the optimal value.

In conclusion, ChrF offers a valuable means of appraising machine translation quality, especially for languages characterized by intricate morphology. Its focus on character-level matches affords a more nuanced evaluation of semantic equivalence compared to word-level metrics. Nonetheless, it is essential to acknowledge ChrF's limitations, such as its potential oversight of fluency and the impact of n-gram selection. By integrating ChrF with other evaluation metrics, researchers and developers can attain a more comprehensive understanding of machine translation quality.

2.6.2.3 COMET

COMET score [51] employs a supervised learning approach, trained on datasets where human evaluators have assessed machine translations across various quality dimensions beyond literal similarity. These encompass fluency, semantic adequacy (maintaining original meaning), and even stylistic elements or sentiment. By analyzing these human assessments, COMET learns to recognize the characteristics of high-quality translations.

Once trained, COMET can assess new machine translations by analyzing the translated text against the source text. By detecting the presence of qualities deemed important by human evaluators, COMET assigns a score to the translation. A higher COMET score indicates a translation more likely to be perceived as high-quality by humans.

COMET represents a significant advancement beyond traditional metrics, such as the BLEU score, which focus solely on n-gram overlap. By integrating insights from human evaluation, COMET offers a more holistic evaluation of translation quality. This positions it as a valuable tool for researchers and developers striving to refine machine translation systems toward producing translations equivalent to those generated by humans. However, it is important to recognize that COMET is still in development. The quality of its evaluation depends on the training data utilized, and any biases or limitations within this data may be reflected in the assigned scores. Furthermore, like all automated metrics, COMET cannot perfectly replicate the complexities of human judgment.

In summary, COMET marks a significant advancement in machine translation evaluation. By incorporating insights from human evaluation, it provides a more nuanced assessment of translation quality.

2.7 Related Work

The main focus of AMT (Arabic Machine Translation) research initially was on translating from Modern Standard Arabic (MSA) to English, with significantly less emphasis on the reverse direction, from English to Arabic, and even fewer efforts dedicated to translations between Arabic and other languages. The first direct-based machine translation system from English to Arabic was developed by Weidner Communication Inc. In 1990, Apptek introduced ArabTrans MT, a tool for translating from English to Arabic. Additionally, products like Al-Mutarjim Al-Araby, Al-Alamiyah, and Al-Nakheel were capable of translating between French and Arabic as well as English and Arabic. A critical challenge in using the Interlingua approach for AMT is the construction of representations that resolve ambiguity and accurately reflect the semantic structure of the language. There have been limited studies that have developed and assessed models using this method. Table 2.5 summarizes the surveyed AMT research studies.

Year	Research	SL-TL[a]	Method	Score
		Linguistic-based AMT researches		
Direct-based AMT				
2005	Al-Taani & Hailat [52]	En→Ar	-	57.3%
2007	Ittycheriah & Roukos [53]	Ar→En	Word alignment	Bl 51.27
Rule-based AMT				
1995	Mankai & Mili [54]	Ar→En/Fr	-	
2008	Salem et al. [55]	Ar→En	-	
2008	Nguyen & Vogel [56]	Ar→En	-	56.03
2008	Samy &	Ar-Sp-En	-	
	González-Ledesma [57]		-	
2008	AbuShuqier & Sembok [58]	En→Ar	-	96.1%
2009	Elming & Habash [59]	En→Ar	-	
2009	Besançon et al. [60]	En/Fr→Ar	-	
2012	Salloum & Habash [61]	DA→MSA	-	
2020	Sghaier & Zrigui [62]	DA→MSA	-	Bl 55.22
Interlingua-based AMT				
2002	Soudi et al. [63]	En→Ar	-	
2006	Shaalan et al. [64]	En→Ar	-	
2008	Bouillon et al. [65]	Jp↔Ar	-	
2014	Al Ansary [66]	En→Ar	Universal Networking Language	
Transfer-based AMT				
2002	Attia [67]	En→Ar	Agreement features	
2004	Shaalan et al. [68]	En→Ar	-	92%

Year	Research	SL-TL[a]	Method	Score
2010	Shirko et al. [69]	Ar→En	-	94.6%
2010	Shaalan et al. [70]	En↔Ar	-	0.450/0.458
2011	Hatem et al. [71]	En→Ar	Morphological analysis	
	Corpus-based AMT researches			
	Example-based AMT			
2002	Guidere [72]	Fr-Ar	-	
2011	Bar & Dershowitz [73]	Ar→En	Verb paraphrases	23.98
2012	Bar & Dershowitz [74]	Ar→En	Semantic equivalents	
2012	Cavalli-Sforza & Phillips [75]	Ar→En	Morphological analysis	
2014	El-Shishtawy & El-Sammak [76]	En→Ar	Template-based syntactic matching	
	Statistical-based AMT			
2006	Hasan et al. [77]	Ar→Fr	Pre-/Post-processing	40.8%
2007	Diab et al. [78]	Ar→En	Pre-/Post-processing	45.38%
2007	Sarikaya & Deng [79]	En→Ar	POS tag[b]/CDW[d]	+0.3
2009	Badr et al. [80]	En→Ar	Syntactic phrase reordering	Bl 32.46
2009	Habash & Hu [81]	Ar→En→Cn	MT Evaluation	Bl +1.1
2010	Carpuat et al. [82]	Ar→En	Word reordering	Bl 51.70
2010	Ghurab et al. [83]	Ar↔Cn	-	0.805/0.696
2010	Bisazza and Federico [84]	Ar→En	Word reordering	48.96
2017	Durrani et al. [85]	Ar↔En	MT Evaluation	Bl +4
2017	Mallek et al. [86]	Ar↔En	Pre- processing	Bl 10.98
2017	Ebrahim et al. [87]	En→Ar	MWE Detection	19.31/19.22
2019	Aqlan et al. [88]	Ar→Cn	Morpho/Vocab/POS tag	Bl 19.40
	Hybrid AMT			
2004	Alsharaf et al. [89]	Fr→Ar	Direct+Transfer+Pivot+SMT	
2008	Toutanova et al. [90]	En→Ar	Inflection prediction models + SMT	Bl 36.92
2008	Hatem & Nassar [91]	En→Ar	Rule-based + Example-based	68%
2009	Matusov et al. [92]	Ar→En	Multi-engine (5 MT systems)	
2009	Habash et al. [93]	Ar→En	SMT + Rule-based	Bl 0.4162
2010	Al Dam & Guessoum [94]	En→Ar	Transfer-based + ANN	56%
2010	Sawaf [95]	DA→MSA	Rule-based+SMT	42.1%
2011	Alawneh & Sembok [96]	En→Ar	Rule-based + Example-based	
2012	Shaalan & Hany [97]	Ar→En	Rule-based + Example-based	WER 88%
2014	Akeel & Mishra [98]	En→Ar	Rule-based + ANN	0.6
2015	Mohamed & Sadat [99]	Ar→Fr	Morphological rule + SMT	34.7%

Year	Research	SL-TL[a]	Method	Score
2015	Zantout & Guessoum [100]	Ar→En	Transfer-based + ANN	64.50%

[a]Source language-Target language [b]Part Of Speech Tagging

[c]Context Dependent Words

Table 2.5: Linguistic-, Corpus-based and Hybrid AMT researches

2.8 Summary

In conclusion, this chapter has delved into the multifaceted challenges and intricacies surrounding Arabic language processing, with a specific focus on machine translation (MT). We have explored the intricate aspects of Arabic, ranging from its morphology and syntax to its diverse array of dialects, all of which present formidable hurdles for natural language processing (NLP) and MT systems alike. Additionally, we have scrutinized the influence of social media on Arabic NLP and MT, taking into account the unique linguistic characteristics and informal language usage prevalent on online platforms.

Our examination has encompassed both linguistic and corpus-based approaches to Arabic MT, underscoring the necessity of tailored resources and datasets to capture the language's nuances effectively. Moreover, we have delved into various MT evaluation methods, encompassing both human and automatic assessment techniques.

Lastly, we have surveyed the related work in the field, providing insights into recent advancements, methodologies, and the challenges encountered in Arabic language MT research, thus offering valuable perspectives on the current landscape and future trajectories of the field.

Chapter 3

Neural Networks and Machine Translation

3.1 Overview

In this chapter, we embark on an in-depth exploration of neural networks, pivotal to understanding the intricate workings of machine translation. Beginning with a thorough examination of the foundational feed-forward networks (FFNs), recurrent neural networks (RNNs), long short-term memory networks (LSTMs), and gated recurrent units (GRUs), we lay the groundwork for comprehending the neural machine translation (NMT) process.

We delve into the nuances of training NMT models, and the decoding phase, where trained models generate translations from input sequences. Finally, we conducted a comprehensive survey of recent advancements and methodologies in neural machine translation research, with a focus on exploring related works specifically dedicated to the Arabic language context.

3.2 Neural Networks

Neural networks are the foundational computational mechanisms driving Neural Machine Translation, employing complex algorithms to mimic human language understanding and generation. At the core of these networks is the node, or processing unit, which functions similarly to neurons in the human brain. Each node is designed to receive inputs—typically a vector of real-valued numbers—process these inputs through a series of mathematical operations, including weighted sums and activation functions, and then produce an output that can be passed on to subsequent layers or nodes in the network. The following subsections will delve into the different types of neural networks, beginning with the simplest form: the feed-forward neural network.

3.2.1 Feed-Forward Neural Networks

A feed-forward neural network is a basic type of neural network where information moves in one direction from the input to the output layer without any loops or cycles back to previous layers. Outputs from each node are forwarded upwards to the subsequent layer without any feedback to lower levels. An illustration, referred to as Figure 3.1, shows a FFN consisting of two layers. These networks are composed of three distinct types of nodes: input, hidden, and output. The input nodes receive individual scalar values, and notably, among these input nodes is a constant bias node, labeled x_0, which is always set to the value of 1. In the hidden layer, nodes calculate a weighted sum of their inputs and then apply a nonlinear function to this sum to determine their output. In a standard neural network structure, every

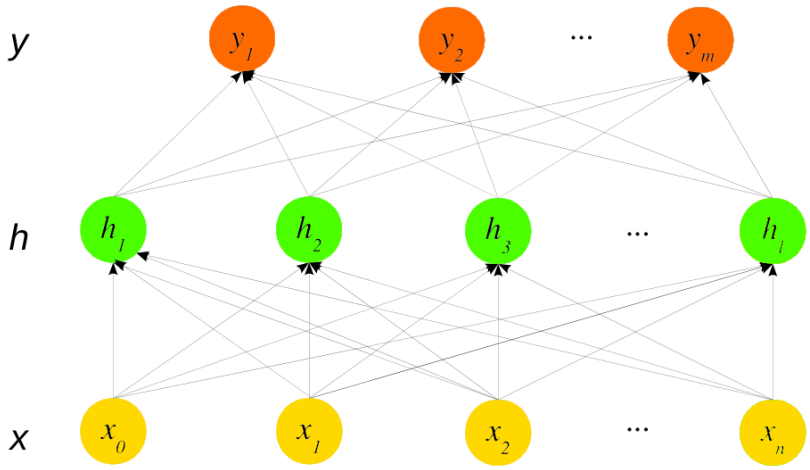

Figure 3.1: A feed-forward network consisting of two layers includes an input layer x, a hidden layer h, and an output layer y.

layer is comprehensively interconnected, allowing nodes to receive inputs from all nodes in the preceding layer. These networks are characterized by their depth, with numerous layers contributing to their complexity. Hidden nodes within these layers act as automatic feature detectors, eliminating the need for manual feature identification by learning to recognize relevant patterns in the input data through training. Each hidden node is defined by its parameters, including a weight vector and a bias term. The parameters for the entire hidden layer are aggregated into a weight matrix U, where each weight u_{jk} in U corresponds to the connection strength from the k^{th} input node x_k to the j^{th} hidden node h_j, and a bias vector that applies to the entire layer.

A notable feature of neural networks is the incorporation of non-linear activation functions, with the rectified linear unit ($ReLU$) standing out for its simplicity and widespread usage. $ReLU$ is particularly efficient to compute, as depicted in Figure 3.2 (a). This function

returns the value of z when z is greater than 0 and returns 0 when z is less or equal to 0. The *sigmoid* (or logistic) function, another frequently utilized activation function, has been

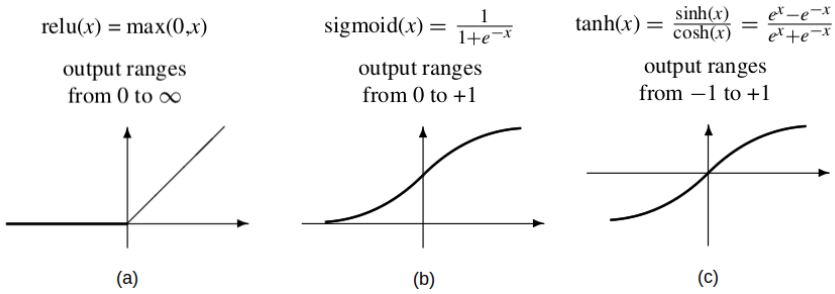

Figure 3.2: Common activation functions utilized in neural networks.

illustrated in Figure 3.2 (b). An activation function similar to the sigmoid but more widely employed is the hyperbolic tangent (*tanh*), depicted in Figure 3.2 (c).

The computation of the hidden layer in the basic feed-forward network can be performed efficiently using straightforward matrix operations, involving three main steps:

- The multiplication of the weight matrix by the input vector x,
- Addition of the bias vector, and
- Application of the activation function f, such as *ReLU*, *sigmoid*, or *tanh*.

Hence, the neural network depicted in Figure 3.1 can be expressed mathematically as follows:

- A set of input nodes represented by the vector $x = (x_1, x_2, x_3, ..., x_n)^T$;
- A set of hidden nodes represented by the vector $h = (h_1, h_2, h_3, ..., h_l)^T$;
- A set of output nodes represented by the vector $y = (y_1, y_2, y_3, ..., y_m)^T$;
- A weight matrix connecting input nodes to hidden nodes denoted by $U = u_{jk}$;
- A weight matrix connecting hidden nodes to output nodes denoted by $W = w_{ij}$.

The computation of the hidden layer output, represented by the vector h, is performed according to Equation 3.1, incorporating the activation function f.

$$h_j = f(\sum_k u_{jk} x_k) \tag{3.1}$$

The value h obtained serves as a representation of the input. Subsequently, the output layer's function is to utilize this revised representation h and calculate a conclusive output, as described in Equation 3.2.

$$y_i = \sum_j w_{ij} h_j \tag{3.2}$$

In numerous instances, this output may initially be a real-valued number; however, it is often transformed into a probability distribution using a $softmax$ function. For a vector y with a dimensionality of d, the $softmax$ function is defined as shown in Equation 3.3, where $1 \leq i \leq d$.

$$softmax(y_i) = \frac{e^{y_i}}{\sum_j e^{y_j}} \tag{3.3}$$

The $softmax$ function operates on a vector $y = [y_1, y_2, y_3, ..., y_m]$ of arbitrary values, transforming them into a probability distribution where each value falls within the range of (0, 1) and the sum of all values equals one.

3.2.2 Recurrent Neural Networks

A neural network containing a cycle inside its network connections is called a recurrent neural network (RNN). Its preceding outputs directly or indirectly influence the value of a node. Figure 3.3 illustrates the structure of a simple RNN based on [101]. Similar to conventional FFNs, the values for a layer of hidden nodes are calculated by multiplying an input vector representing the current input, x, by a weight matrix and then passing the result through a non-linear activation function. The associated output, y, is then determined using the hidden layer, which comprises the hidden nodes. The context layer is where it differs from a FFN the most. It keeps the previous values and sends them to the appropriate nodes in the hidden layer. This layer uses the hidden layer's value from the previous time step as input to the computation at the hidden layer.

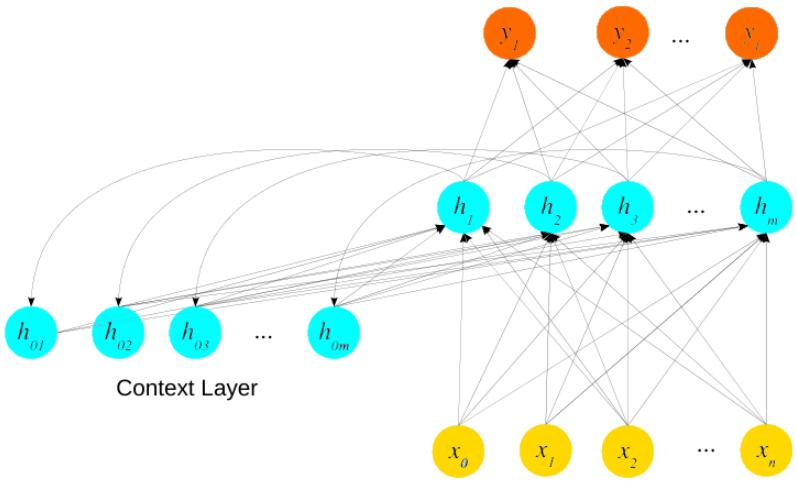

Figure 3.3: A basic recurrent neural network.

3.2.2.1 Stacked Recurrent Neural Networks

Stacked recurrent neural networks (RNNs) consist of multiple layers, where the output of one layer serves as the input to the next layer. Information flows from lower layers to higher layers, and the final output is generated by the top layer. Stacked RNNs have shown improved performance in neural MT compared to single-layer networks [102]. This improvement can be attributed to the network's ability to create representations at different levels of abstraction across the layers. However, the optimal number of stacked RNNs depends on the availability of training data. While abundant data can lead to better generalization, limited data may result in overfitting [14]. Moreover, increasing the number of stacked layers escalates the training costs.

3.2.2.2 Bidirectional Recurrent Neural Networks

In recurrent networks, the hidden state at any given time encapsulates all information about the sequence up to that point, serving as the left context for the current input. In neural machine translation (NMT), having access to the entire input sequence simultaneously suggests the importance of utilizing the context to the right of the input. Training an RNN in reverse on the input sequence is one approach to retain this information. Combining the forward and backward networks results in a bidirectional RNN, where two independent RNNs process input from start to finish and finish to start, respectively [103]. The outputs of both networks are then merged to create a unified representation incorporating both left and right input contexts at each time step. Methods such as concatenation, element-wise operations, or averaging can be employed to combine the outputs of the forward and backward passes. This ensures that the information on both sides of the current input is captured in the output at each time step. Figure 3.4 illustrates the conventional RNN structure alongside the BRNN architecture over a timeline. In the RNN, only a forward hidden layer is present, whereas in the BRNN, both forward and backward hidden layers are depicted.

3.2.3 Long Short-Term Memory

One of the main limitations of RNNs in machine translation is their difficulty in handling long-distance dependencies. Distant words play a crucial role in translation tasks, as demonstrated in the following example:

> The lengthy *sentence* composed of numerous words spanning multiple lines still *poses* significant challenges for MT.

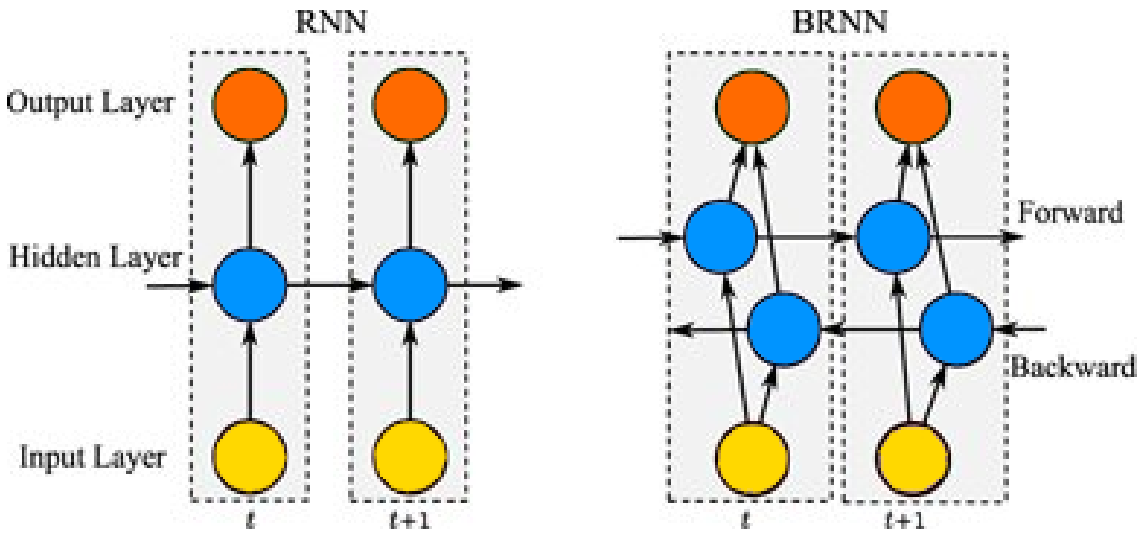

Figure 3.4: Architecture of RNN and BRNN shown over a period of time.

In this example, the verb *poses* depends on the subject *sentence*, which is separated by a long subordinate clause.

While RNNs can access preceding sequences, the knowledge retained in hidden states tends to be relatively localized. Consequently, more complex network structures have been developed to address the difficulty of maintaining context over extended periods. It is essential for the network to discard irrelevant information while retaining data crucial for future decision-making. Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRUs) are the predominant approaches employed to achieve this objective.

LSTM networks, introduced by [104], address the challenge of managing context by dividing it into two sub-problems: removing unnecessary information and incorporating data crucial for future decisions. Unlike incorporating a fixed strategy into the architecture, the key lies in learning to handle context dynamically. LSTMs accomplish this by expanding the architecture with an explicit context layer and employing specialized neural units with gates to control the flow of information within the network layers. These gates utilize additional weights to sequentially process input, previous hidden layers, and previous context layers.

3.2.4 Gated Recurrent Units

GRU, akin to LSTM but with fewer parameters, offers the advantage of reduced training costs by requiring fewer parameters. This reduction in parameters is achieved by eliminating the need for a separate context vector and reducing the number of gates, thereby alleviating the computational burden compared to LSTM. Both GRU and LSTM employ sigmoid functions in their gates to either allow information with values close to one to pass through or block information with values close to zero.

In contrast to simple FFNs, LSTMs and GRUs feature more complex neural nodes, as depicted in Figure 3.5, with inputs and outputs connected to each node type. The additional complexity within LSTM and GRU nodes is primarily contained within the nodes themselves. The only external complexity introduced by LSTM, beyond the basic recurrent node, is the availability of the other context vector as an input and output. Conversely, GRU nodes exhibit similar input and output architecture to the simple recurrent node.

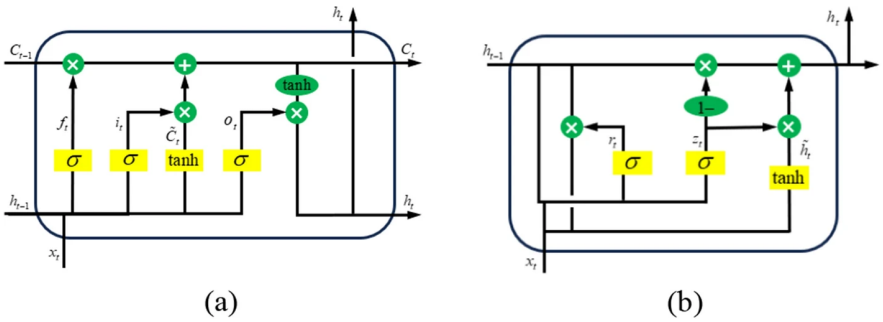

Figure 3.5: Diagram illustrating the network structure of LSTM in the first panel (a) and GRU in the second panel (b)

3.3 Neural Machine Translation

The standard architecture for Neural Machine Translation (NMT) is the encoder-decoder network. This architecture operates on the fundamental principle of employing an encoder network to convert a sequence of words from the source language sentence into a contextual representation. Subsequently, a decoder utilizes this representation to generate an output sequence, essentially providing a plausible translation into the target language. Figure 3.6 illustrates the encoder-decoder architecture at its most abstract level, comprising three key components: an encoder, context, and decoder. The encoder takes in a source language sentence as an input sequence of words, $x_1, x_2, ..., x_m$, and produces a corresponding sequence of contextualized representations known as a context vector. This context vector captures the essence of the input and serves as input to the decoder. Finally, the decoder utilizes the context vector to generate the most probable translation as a sequence of words, $y_1, y_2, ..., y_n$. Encoders and decoders can be implemented using either RNNs or Transformers.

Figure 2.7 illustrates a simplified version of the RNN-based encoder-decoder architecture. The encoder processes the input sequence x with the objective of generating a representation of the input. This representation is captured in the final hidden state of the encoder, represented as h_n^e. Subsequently, this context representation, denoted as c, is transferred to the decoder. Upon receiving this state, the decoder utilizes it to initialize its initial hidden state.

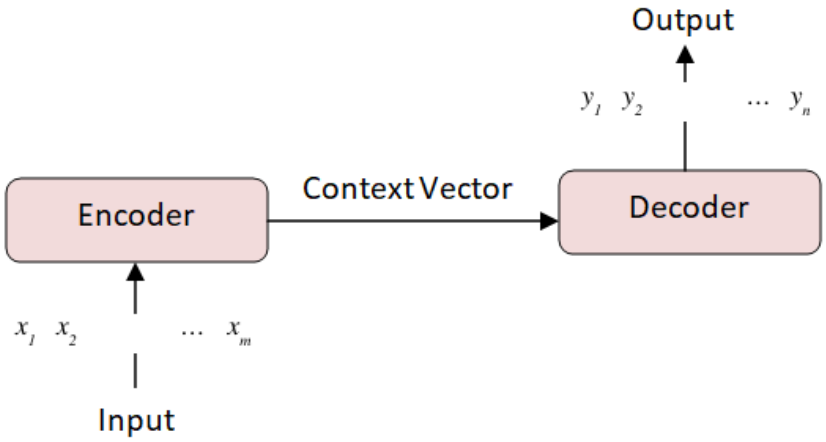

Figure 3.6: Basic Encoder-Decoder architecture

The first decoder cell employs c as its initial hidden state, h_0^d, and proceeds to generate a sequence of outputs iteratively, one element at a time, until an end-of-word marker, </s>, is produced. Consequently, each hidden state is dependent on the preceding hidden state and the output generated in the preceding state. The embedding layer is composed of word embeddings, which adhere to the notion that semantically related words in similar contexts should possess similar representations. Ultimately, the output y at each time step involves a $softmax$ computation over the vocabulary, V. The most probable output at each time step can be determined by computing the $argmax$ over the $softmax$ output as per Equation 3.4. Although Figure 3.7 illustrates a single network layer, stacked and bi-directional networks are typically employed for both the encoder and decoder in practical implementations.

$$\hat{y}_t = argmax_{w \in V} P(w|x, y_1, ..., y_{t-1}) \tag{3.4}$$

A significant drawback of this architecture lies in its inability to evenly represent information from the beginning of a sentence, particularly evident in lengthy sentences. To address this issue, the attention mechanism emerges as a solution. It enables the decoder to access information from all the hidden states of the encoder, rather than solely relying on the last hidden state. The concept behind the attention mechanism is to create a singular context vector, denoted as c, by computing a weighted sum of all the encoder's hidden states. These weights concentrate on a specific segment of the source text pertinent to the token generated by the decoder. Notably, the context vector produced by the attention mechanism is dynamic, varying for each decoded token. By introducing the attention mechanism, the static context vector is replaced with c_i a dynamically derived counterpart from the encoder's hidden states at each decoding step i, ensuring comprehensive consideration of all encoder states.

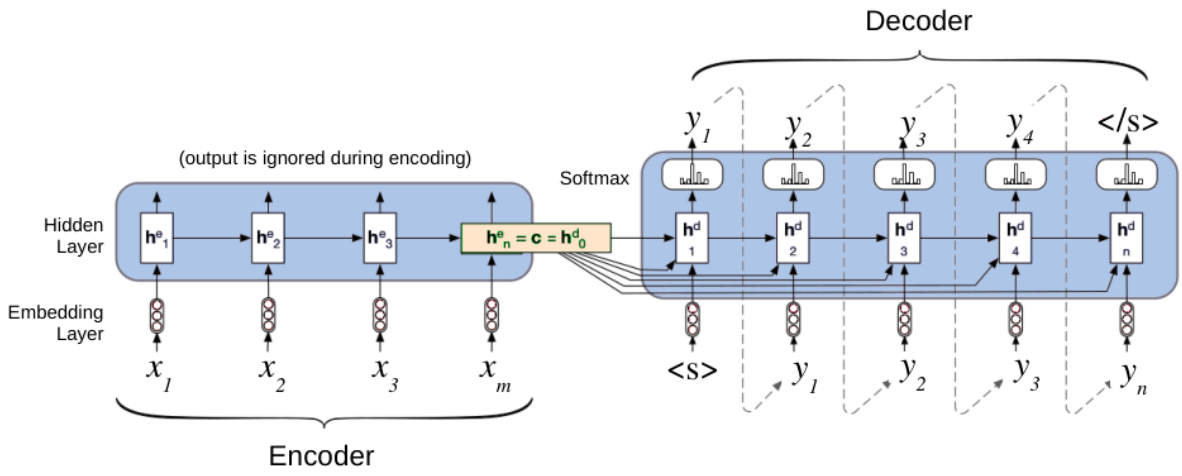

Figure 3.7: The basic RNN-based encoder-decoder architecture.

3.3.1 Training Neural Machine Translation Models

Neural Machine Translation (NMT) operates within a supervised machine learning framework, where the correct output, denoted as y, is known for each input observation, x. The system's output, represented as $\hat{y}$, serves as an estimate of the actual y. During the training phase, the objective is to adjust the parameters within each layer so that $\hat{y}$ closely resembles y for every training instance. To accomplish this, a loss function is employed to quantify the disparity between the system's output and the actual output. The optimization process, typically facilitated by the gradient descent algorithm, seeks to minimize this loss function by adjusting the parameters. Gradient descent utilizes the gradient of the loss function, computed as the partial derivative of the loss function with respect to each parameter, to iteratively update the model's parameters. However, in the realm of neural networks, characterized by numerous layers and millions of parameters, computing the partial derivative of a weight in one layer concerning the loss associated with another layer necessitates employing techniques such as error back-propagation or reverse differentiation. These methodologies are crucial for efficiently navigating the complex landscape of neural network optimization.

Moreover, NMT models undergo end-to-end training, where each training instance comprises a pair of sentences, one in the source language and the other in the target language. These sentence pairs are concatenated with a designated separator token, <s>, to form the training data. Typically, this training data is sourced from established datasets containing aligned pairs of sentences, known as parallel corpora. Optimization within the NMT framework is characterized by a non-convex problem landscape. Nevertheless, there exist several best practices for effectively training NMT models. For instance, it is advisable to initialize

the model weights with small random numbers and to employ random seeds for reproducibility. Additionally, normalizing the input values to have zero mean and unit variance can be beneficial in enhancing training stability and convergence.

NMT model training involves multiple epochs, representing complete iterations over the training data. Typically, as training progresses, the error on the training set steadily decreases. However, a key challenge arises when the model starts to overfit, meaning it memorizes the training data excessively without generalizing well to unseen examples. To diagnose overfitting, a separate set of examples known as the development (or validation) set is used, which is not involved in the training process. Monitoring the error on this development set over the course of training reveals a point where the error begins to increase, indicating overfitting. In theory, training should halt when the minimum error on the development set is reached. However, in practice, this is more nuanced for NMT models due to the non-deterministic nature of training and the lack of clear convergence or overfitting points, particularly with sizable datasets. While some studies suggest stopping criteria based on approximate training durations or a fixed number of epochs, most NMT research does not specify precise stopping conditions. To mitigate overfitting, various regularization techniques are employed, such as dropout, which randomly disables some nodes and connections during training.

In addition to regularization, hyperparameter tuning plays a crucial role in NMT architecture design. These hyperparameters, which include the learning rate, mini-batch size, number of layers, hidden nodes per layer, and activation functions, are chosen by the architect to optimize model performance.

3.3.2 Decoding

The decoding (inference) algorithm employed to generate translations (as illustrated in the Figure 3.7) faces a challenge. Opting for the single most probable word at each step implies a 1-best greedy search, where a greedy algorithm makes locally optimal decisions. However, there are instances where following a sequence of words leads to the realization that an earlier mistake was made. In such cases, the best sequence may initially comprise less probable words that are refined by subsequent words in the context of the entire output. For instance, when translating an idiomatic expression, the initial words chosen may seem unusual (e.g., "*take your hat off*" for "*respect, admire, or congratulate someone*").

In NMT decoding, the predominant method employed is known as beam search. Unlike selecting the best word at each step, beam search maintains k possible words, where k denotes the beam size or beam width. Initially, a $softmax$ is computed over the entire vocabulary to assign probabilities to each word. From this $softmax$ output, the k-best candidates are

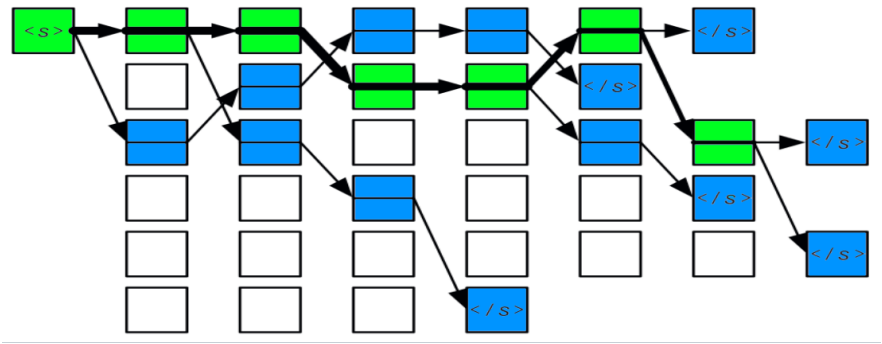

Figure 3.8: Beam search decoding with a beam size of six.

selected, forming hypotheses. Each hypothesis represents an initial output sequence along with its probability. These top k hypotheses are progressively expanded by passing them through different decoders in subsequent steps. At each decoder, a $softmax$ is generated over the entire vocabulary to determine the next word for each hypothesis. Subsequently, each of the $k * V$ potential sequences is scored based on the product of the current word's probability and the path leading to it. These hypotheses are then pruned down to the k best ones to ensure there are never more than k hypotheses at any given time. This iterative process continues until a complete translation candidate marked by "</s>" is generated, signifying the end of a hypothesis. The finished hypothesis is then removed, and the beam size is reduced by one. This search persists until the beam size diminishes to zero, resulting in k hypotheses. The best translation is determined from the complete hypothesis with the highest score. Figure 3.8 illustrates this process using a beam size of six. When evaluating the top paths, we assess each based on the product of their word prediction probabilities. In practical applications, we tend to achieve improved outcomes by normalizing the score according to the length of the translation produced, which involves dividing the score by the number of words. This normalization process occurs after the search is finalized.

3.4 Related Work

In the recent years, Research in AMT has predominantly centered on Neural MT, targeting translation from Arabic to English and various other languages. Neural machine translation, in particular, has become a compelling substitute for phrase-based Statistical Machine Translation (SMT), especially for the Arabic language. Numerous studies are currently focusing on the translation between Dialectal Arabic and Modern Standard Arabic (MSA) in low-resource settings, both from Dialectal Arabic to MSA and vice versa. Table 3.1 summarizes the surveyed Neural-based AMT research studies.

Table 3.1: Neural-based AMT researches

Year	Research	SL-TL[a]	Method	Score
Neural-based AMT				
2016	Almahairi et al. [105]	Ar↔En	MT Evaluation	49.7/33.62
2016	Guzmán et al. [106]	En→Ar	Morpho-syntactic analysis	+75%
2017	Belinkov et al. [107]	Ar↔En	Morpho/Vocab/Factored NMT	Bl 28.42
2017	Durrani et al. [85]	Ar↔En	MT Evaluation	Bl +4
2017	Choi et al. [108]	Kor→Ar	Corpus extension	Bl 27.07
2018	Almansor & Al-Ani [22]	Ar→En	Low-resource	Bl +10
2018	Shapiro & Duh [109]	Ar→En	Morpho/Vocab/Factored NMT	Bl 29.10
2018	Alrajeh [110]	Ar→En	MT Evaluation	Bl +13
2018	Alkhatib & Shaalan [111]	Ar↔En	Paraphrasing model	86.9%/94.1%
2018	Baniata et al. [112]	DA→MSA	Multitask Learning	0.41/0.30
2019	Aqlan et al. [113]	Ar→Cn	Pre-/Post-processing	Bl 24.66
2019	Oudah et al. [114]	Ar→En	Pre-/Post-processing	55.64/53.54
2019	Hadj Ameur et al. [115]	Ar↔En	Pre-/Post-processing	29.76/20.41
2019	Gashaw & Shashirekha [116]	Amh→Ar	MT Evaluation	Bl 12
2020	Ji et al. [117]	→Ar	Multilingual/Low-resource	25.49
2021	Bensalah et al. [118]	Ar→En	Pre-processing/MT Evaluation	41.87%
2021	Berrichi & Mazroui [119]	Ar↔En	Morpho/Vocab/Factored NMT	Bl 33.02
2021	Moukafih et al. [120]	DA↔Ar	Multi-Task learning/MT Eval	Bl 35.06
2021	Nagoudi et al. [121]	DA→En	Pre-processing/MT Evaluation	Bl 25.72
2021	Stergiadis et al. [122]	Ar→En	Multidimensional Tagging	Bl 50.84
2022	Bensalah et al. [123]	Ar→En	MT Evaluation	Bl 0.575
2022	Slim et al. [124]	DA→MSA	Transductive transfer learning	Bl 35.87
2022	Gaser et al. [125]	DA→En	Segmentation	chrF2 51.3
2022	Nagoudi et al. [126]	20 lang→Ar	Multi-lingual model	Bl 32.07
2022	Hameed et al. [127]	Ru→Ar	MT Evaluation	51.57%
2022	Baniata et al. [128]	DA→Ar	RPE[b]+sub-word	Bl 66.87

[a]Source language-Target language [b]Reverse Positional Encoding

3.5 Summary

In this chapter, we embarked on a comprehensive exploration of neural networks, particularly focusing on feedforward networks (FFN), recurrent neural networks (RNN), long short-term memory networks (LSTM), and gated recurrent units (GRU). These architectures lay the foundation for understanding the mechanisms behind neural machine translation (NMT).

We delved into the intricacies of training NMT modelsand the decoding phase, where the trained models generate translations from input sequences. Finally, we delved into related works in the field, dedicating special attention to advancements, methodologies, and challenges specific to neural machine translation in the context of the Arabic language.

Chapter 4

Large Language Models

4.1 Overview

This chapter provides an in-depth exploration into the transformative landscape of large language models (LLMs) and their pivotal role in revolutionizing natural language processing (NLP) and machine translation (MT).

We begin by delving into the foundational architecture of LLMs, notably Transformers, BERT, and GPT, which have reshaped the understanding of linguistic patterns and relationships. Within this context, we examine how LLMs, particularly BERT and GPT, have evolved to address bidirectional context understanding and generative text tasks, respectively.

Furthermore, we explore the application of LLMs in MT, showcasing variants like BERT, T5, and ChatGPT, which demonstrate promising potential in overcoming language barriers and enhancing translation accuracy.

Following this exploration, we present a comprehensive section on related work, offering insights into the latest research and advancements in the field of LLM-based MT.

4.2 Transformers

While LSTMs and GRUs alleviate the issue of losing distant information inherent in simple RNNs, they are unable to leverage parallel computing resources due to their sequential nature. This limitation is addressed by Transformers, as introduced by [1]. Transformers revolutionize sequence processing by completely replacing RNNs. They operate by mapping sequences of input vectors $(x_1, x_2, ..., x_m)$ to sequences of output vectors $(y_1, y_2, ..., y_n)$ through stacks of network layers featuring customized connections of basic feed-forward networks and self-attentions. Unlike RNNs, Transformers enable the extraction and utilization of information from broader contexts through self-attention mechanisms, without the need for recurrent

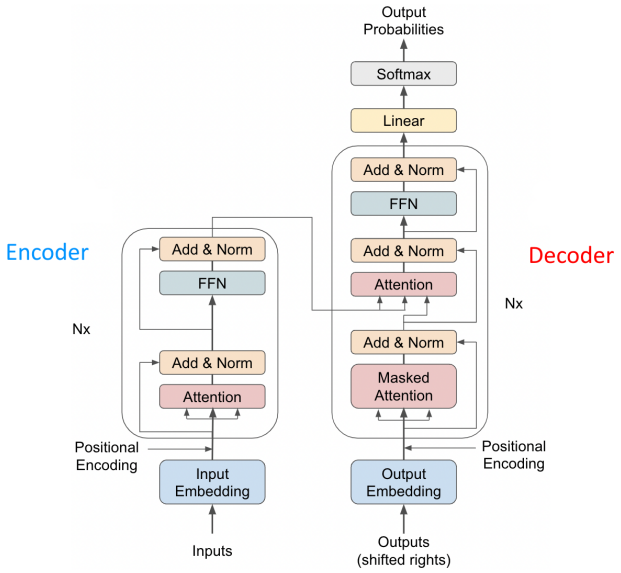

Figure 4.1: Transformer architecture adapted from [1]

intermediary connections.

The Transformer, as depicted in Figure 4.1, is a sequence-to-sequence (Seq2Seq) network that captures sequential information through stacked self-attention and cross-attention layers. The output O of each attention sub-layer is calculated using scaled multiplicative formulations, as defined by:

$$A = (QW^Q)(KW^K)^T/\sqrt{d}; \quad Att(Q,K,V) = softmax(A)(VW^V) \tag{4.1}$$

$$O = Att(Q,K,V)W^O \tag{4.2}$$

Where, $Q = (q_1, ..., q_{l_q}) \in \mathbb{R}^{l_q \times d}$,$K = (k_1, ..., k_{l_k}) \in \mathbb{R}^{l_k \times d}$, $V = (v_1, ..., v_{l_k}) \in \mathbb{R}^{l_k \times d}$ represent matrices of query, key, and value vectors respectively. Additionally, W^Q, W^K, W^V, and $W^O \in \mathbb{R}^{d \times d}$ denote the associated trainable weight matrices. A signifies the affinity scores (or attention scores) between queries and keys, while $Att(Q,K,V)$ represents the attention vectors. In practical applications, multi-head attention is utilized rather than single-head attention. Here, the hidden dimension d is divided into h segments, each processed independently through an attention layer before being amalgamated back together. Subsequently, the final output of a Transformer layer is determined as follows:

$$\phi(A,Q) = LN(FFN(LN(O+Q)) + LN(O+Q)) \tag{4.3}$$

Here, ϕ denotes the standard sequential operations of a Transformer layer incorporating layer normalization (LN) and feed-forward (FFN) layers. The feed-forward layers essentially consist of sequences of fully connected layers interspersed with $ReLU$ activation, which are applied to each token's latent vector separately.

Since attention layers are insensitive to order, implying that the arrangement of key sequences doesn't influence the outcome of a specific query, it's crucial to introduce some form of ordering information for the model to grasp temporal characteristics of the input. To address this, the Transformer integrates positional encoding into the word embeddings immediately after the embedding layer and before the initial attention layer. This positional encoding relies on sine (Equation 4.4) and cosine (Equation 4.5) functions and is defined as follows:

$$E(pos, 2i) = \sin(pos/10000^{2i/d}) \tag{4.4}$$

$$E(pos, 2i+1) = \cos(pos/10000^{2i/d}) \tag{4.5}$$

Here, pos denotes the position of the word within the sequence, while i is the dimension.

4.3 Large Language Models

At the forefront of Natural Language Processing (NLP) stands a novel class of models termed Large Language Models (LLMs). These intricate artificial neural networks undergo training on massive volumes of textual data. Diverging from conventional NLP models tailored to specific tasks, LLMs acquire a broader comprehension of linguistic structures and correlations. This versatility enables them to undertake a diverse array of tasks, encompassing sentiment analysis, question answering, text summarization, and even creative writing. Notably, LLMs have significantly propelled the field of machine translation, a critical conduit between languages. Through extensive analysis of translated text datasets, these models can grasp the nuances inherent in different languages, yielding translations that are notably more precise and natural compared to traditional rule-based methodologies. However, the effectiveness of LLMs crucially depends on the caliber and extent of their training data. Greater diversity and comprehensiveness in the dataset enhance the model's understanding of the intricacies of human language. This has led to the exploration of specific LLM architectures such as BERT, mBERT, and GPT, among others, each contributing to the ongoing evolution of NLP and machine translation through their distinctive training methods and capabilities.

4.3.1 BERT

BERT, or Bidirectional Encoder Representations from Transformers, emerged as a groundbreaking pre-training method in the realm of Natural Language Processing (NLP). Unveiled by [129], BERT reshaped the landscape by achieving top-tier performance across a diverse array of NLP tasks. Its standout feature lies in its adeptness at leveraging extensive, unlabeled text corpora during pre-training. Unlike earlier approaches that relied on task-specific

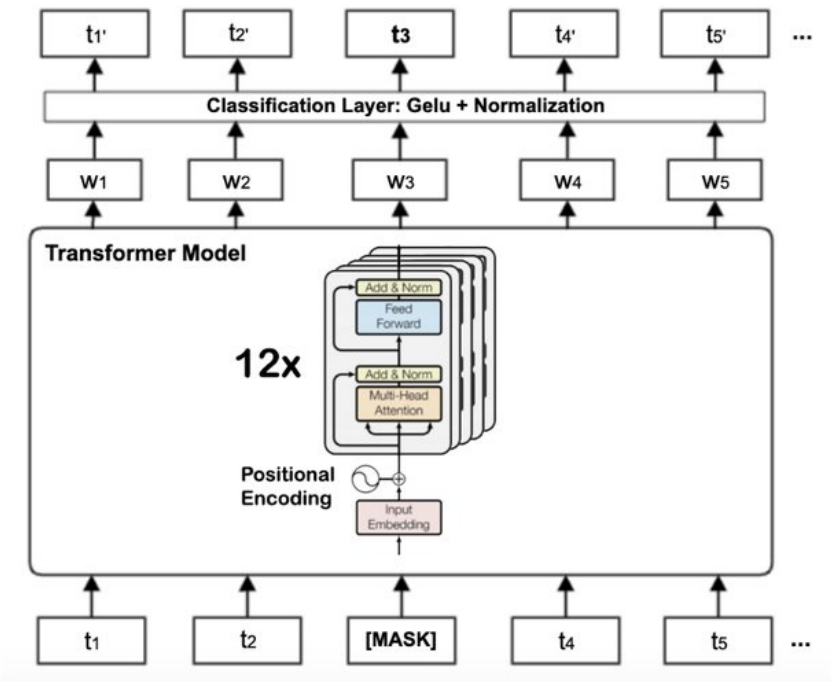

Figure 4.2: BERT base architecture with twelve encoder blocks, adapted from [2].

labeled data, BERT follows a two-phase strategy. Initially, it undergoes pre-training on tasks like masked language modeling and next sentence prediction, exposing it to various linguistic structures and relationships. This equips the model with a robust grasp of word semantics and contextual nuances, forming a strong foundation for fine-tuning on specific tasks. BERT produces embeddings as its output, rather than predicting the next words in a sequence. To utilize these embeddings, additional layers must be added on top, such as those for text classification or question answering tasks.

BERT's architectural backbone relies on the Transformer encoder (see Figure 4.1), a potent neural network architecture introduced by [1]. Unlike traditional sequential models, the Transformer enables parallel processing of entire input sentences, facilitating the capture of long-range dependencies essential for tasks such as sentiment analysis and question answering (see Figure 4.2). Notably, BERT employs a pre-trained encoder model and foregoes the decoder component typically found in sequence-to-sequence models like those used in machine translation. This design choice underscores BERT's focus on comprehending text rather than generating new sequences. A key contributor to BERT's efficacy is the scale of its pre-training data, drawn from vast text corpora like BookCorpus and English Wikipedia.

While BERT's original training was centered on English, subsequent research has explored its applicability to multilingual NLP tasks. mBERT, short for multilingual BERT, extends

the capabilities of BERT to encompass tasks spanning multiple languages. It inherits BERT's architecture but undergoes training on an extensive dataset containing text samples from a staggering 104 languages. Efforts such as mBERT have demonstrated the feasibility of fine-tuning pre-trained models across multiple languages, showcasing BERT's potential for cross-lingual NLP endeavors[130].

4.3.2 GPT

In contrast to BERT's emphasis on pre-training encoders, Generative Pre-trained Transformer (GPT) models offer a compelling alternative for tasks involving text generation. Introduced by OpenAI[1], GPT utilizes a decoder-based architecture derived from the Transformer family. This design enables GPT to excel in generating various forms of creative text, including poems, code, scripts, musical compositions, and even realistic chat dialogues [131].

Unlike BERT, GPT is predominantly trained on a single, extensive dataset consisting of text and code. The initial GPT model, GPT-1, was trained on a dataset compiled from internet sources, while subsequent versions like GPT-2 and GPT-3 have utilized progressively larger and more diverse datasets. Focusing on a single language, typically English, allows GPT to grasp statistical patterns and sequential dependencies within the text, thereby generating text of human-like quality that can often be mistaken for content written by humans.

The fundamental architecture of GPT relies on a Transformer decoder (see Figure 4.1). Unlike encoders that process the entire input sequence simultaneously, decoders handle the input sequence one element (word) at a time. This iterative approach enables GPT to predict the next word in a sequence based on previously generated words and the overall context, making it proficient in tasks like text summarization, machine translation, and creative writing, where generating coherent text is crucial.

In language scenarios, decoders play a crucial role in generating subsequent words, such as in text translation or story generation, where the outputs are words with probabilities. Decoders integrate attention mechanisms, employing them twice in their operation. Initially, during model training, they utilize Masked Multi-Head Attention, where only the initial words of the target sentence are revealed to prevent the model from cheating during learning. This approach resembles the MASK concept introduced in BERT. Subsequently, decoders employ Multi-Head Attention [133], similar to how it is utilized in the encoder. Transformer-based models incorporating both encoders and decoders employ a technique for enhanced efficiency. The output of the encoders serves as input to the decoders, specifically as keys and values. Decoders can then make queries to locate the most relevant keys. This facilitates tasks such as

[1]https://openai.com/

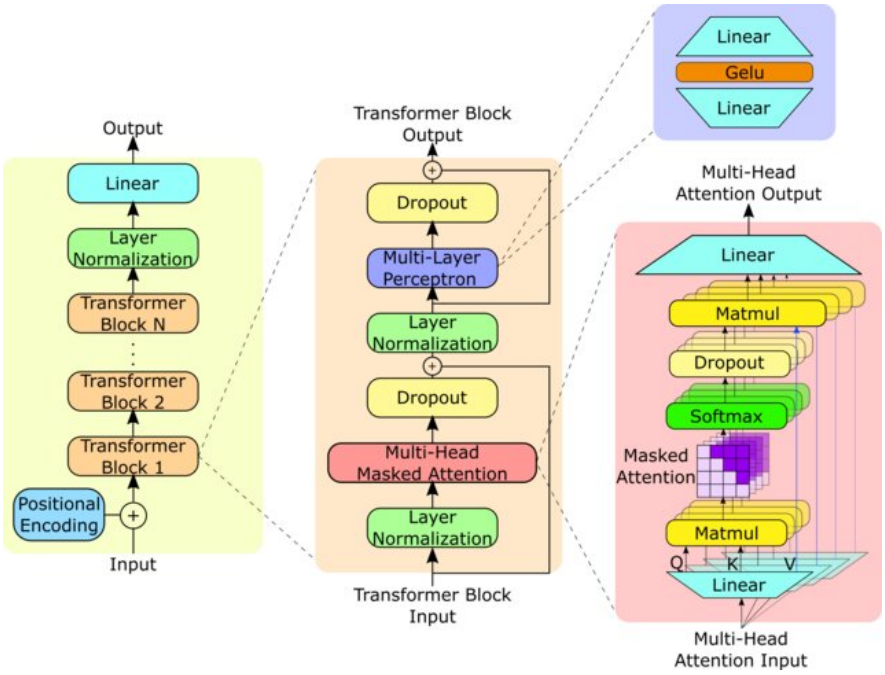

Figure 4.3: GPT-2 model architecture adapted from [132].

comprehending the original sentence's meaning and translating it into other languages, even when the resulting word count and order vary. Figure 4.3 illustrates the structure of GPT-2, consisting of N Transformer decoder blocks. Each block is equipped with multi-head masked attention, multi-layer perceptrons, normalization, and dropout layers. Residual connections link these blocks, enhancing the model's ability to learn from earlier inputs. The multi-head masked attention mechanism employs Q, K, and V vectors for calculating attention scores, effectively capturing and encoding the sequential relationships within the data.

4.4 LLM-based Machine Translation

Unlike conventional rule-based methodologies that depend on predetermined linguistic principles, Large Language Models (LLMs) utilize extensive real-world translation datasets to grasp the intricacies and statistical regularities of language [134]. This enables them to generate translations that are more natural-sounding and precise compared to traditional techniques.

A significant advantage of LLM-based MT lies in its capacity to comprehend context. By analyzing extensive sets of translated texts, LLMs can discern the connections between words, phrases, and the overall context of a sentence. This contextual comprehension empowers them to produce translations that not only adhere to grammatical rules but also capture the intended meaning and nuances of the source language.

Numerous research studies have investigated the efficacy of LLMs in MT. For example, the study conducted by showcases how LLMs can achieve superior performance on diverse language pairs, sometimes outperforming traditional statistical MT approaches.

Nevertheless, LLM-based MT encounters certain challenges. One concern is the potential presence of bias in the training data, which may manifest in the translated output [135]. Translating content into English may enhance multilingual NLP tasks for English-centric LLMs, yet it's not always the best approach, particularly for tasks requiring deep cultural and linguistic comprehension. Using the native language directly often yields better results by capturing cultural nuances more effectively [136]. Additionally, LLMs can pose computational challenges due to their extensive size and complexity, making training and deployment resource-intensive [137]. Despite these hurdles, LLM-based MT offers a promising avenue for seamless cross-lingual communication. As research advances and training datasets expand, we anticipate further progress in this dynamic field.

4.4.1 BART

BART (Bidirectional and Autoregressive Transformer) arises as a potent Large Language Model, differentiating itself from GPT's focus on decoders and BERT's emphasis on encoders. It achieves this by integrating both encoder and decoder elements, enabling it to tackle tasks related to both Natural Language Generation (NLG) and Natural Language Understanding (NLU). This dual structure synergizes the comprehensive understanding capabilities of the encoder with the fluent generative abilities of the decoder.

As highlighted in [138], BART's architecture enables it to deliver superior performance in MT by thoroughly grasping the source language context with its encoder and producing coherent translations with its decoder. The ability to pre-train on extensive multilingual datasets further enhances BART's proficiency in handling a wide array of language pairs, positioning it as an effective solution for overcoming language barriers. Figure 4.4 illustrates the BART model adapted for MT. In this scenario, a supplementary encoder is introduced to substitute word embeddings within BART. This new encoder has the flexibility to utilize a separate vocabulary.

4.4.2 T5

T5, or Text-to-Text Transfer Transformer, emerges as a formidable player in the landscape of LLM-based MT. Unlike models such as BERT, which concentrate on pre-training encoders, or GPT, which prioritize decoders, T5 takes a unified approach. It employs a single Transformer-based architecture that can be fine-tuned for various NLP tasks, including ma-

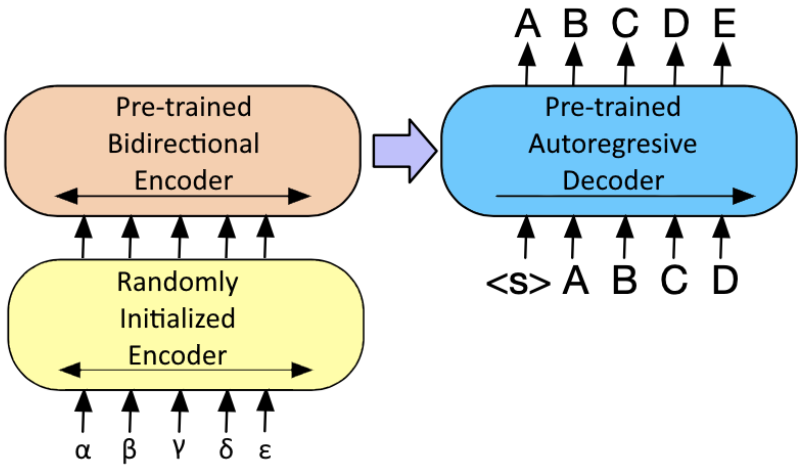

Figure 4.4: Fine tuning BART for machine translation.

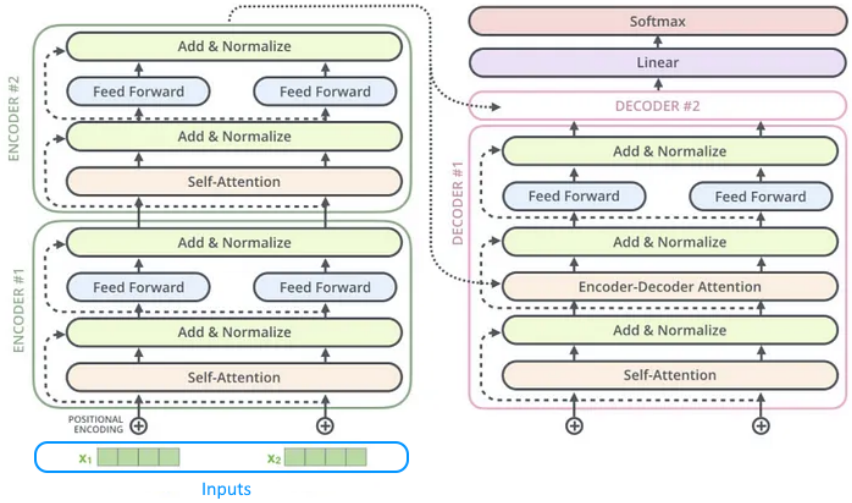

Figure 4.5: T5 model architecture adapted from [3].

chine translation (see Figure 4.5).

The model can be trained on extensive multilingual text and code datasets, enabling it to comprehend the nuances of different languages and perform effective translations between them. Moreover, T5's capability to adjust to diverse NLP tasks equips it to handle a wide range of translation scenarios, from straightforward sentence translations to intricate document summarization with translation components. This versatility positions T5 as a valuable resource for applications necessitating seamless and nuanced language transfer across various domains.

4.4.3 ChatGPT

While ChatGPT demonstrates remarkable proficiency in generating diverse text formats creatively, its utility in MT is circumscribed by certain constraints [139]. Unlike specialized models like BART or T5 tailored explicitly for managing both source and target languages, ChatGPT's fundamental capability resides in its decoder-centric architecture (refer Figure 4.6).

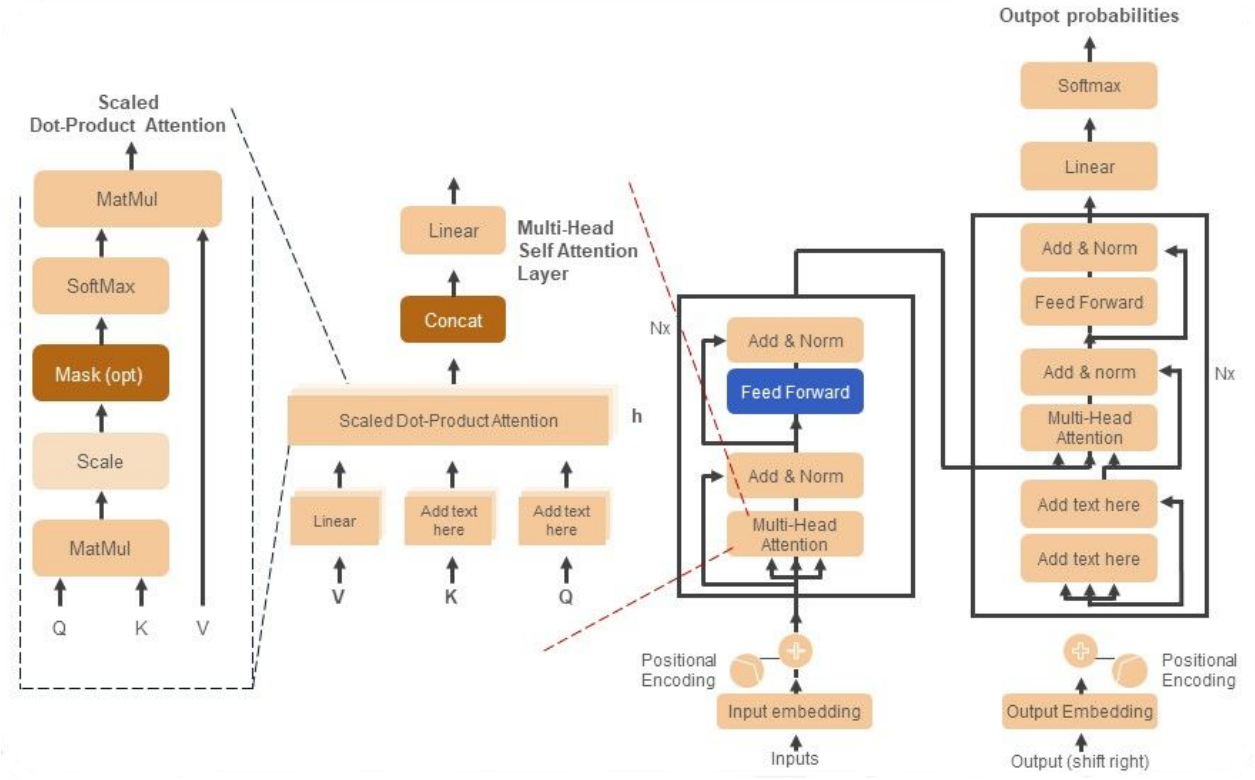

Figure 4.6: Transformer architecture of ChatGPT

This design excels in text generation based on provided prompts or contexts, rendering it suitable for tasks such as creative writing or text summarization. owever, for achieving precise and natural-sounding translations, comprehending the context of the source language is imperative. ChatGPT's primary emphasis on text generation within a single language, typically English (GPT data proportion is only 7% non-English [140]), constrains its capacity to fully comprehend the subtleties and intricacies involved in translating between diverse languages. Therefore, while it can generate text in various languages based on the input it receives, its proficiency in languages other than English may vary depending on factors such as the availability and diversity of training data in those languages (see Figure 4.7).

While some research endeavors explore the potential of adapting ChatGPT for multilingual purposes [141, 142, 143, 140], it has been observed that ChatGPT often generates

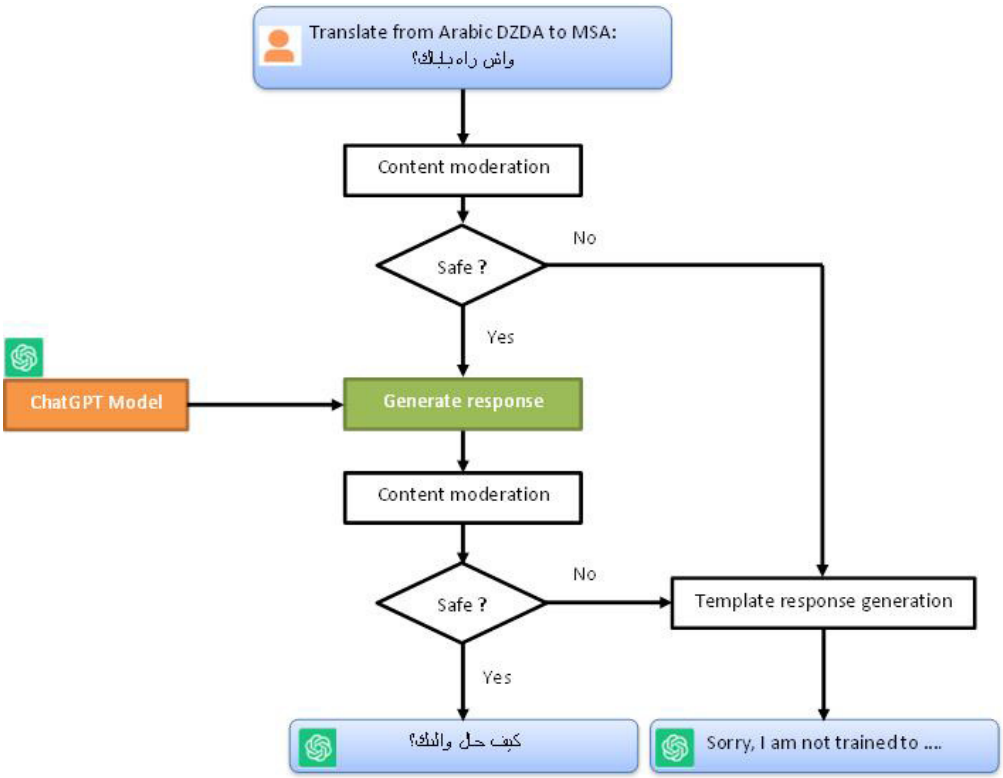

Figure 4.7: A flowchart illustrating the process of how ChatGPT answers a prompt.

inaccurate outputs and hallucinates for non-English-centric machine translation tasks [143]. Its performance generally lags behind models like mBERT or BART, specifically designed for multilingual understanding and translation. Additionally, studies have shown that while ChatGPT can compete with commercial translation products such as Google Translate and Microsoft Translator for high-resource languages, it exhibits limited capabilities for low-resource and distant languages [143, 144, 145, 140].

Recent research studies [141, 146, 147] have underscored the benefits of in-context learning for Large Language Models (LLMs). In-context learning entails embedding input-output examples directly into the input text, serving as prompts, to enhance LLM performance across various tasks, all without necessitating alterations to parameters or architecture. However, the efficacy of in-context learning is contingent upon the architecture of the LLM, as indicated by Olmo et al. [148], with additional examples not consistently leading to significant performance improvements [144].

4.5 Related Work

Brown et al. [141] showcased the considerable benefits of scaling up language models, revealing that larger models notably enhance their few-shot performance across various tasks,

often matching or surpassing the effectiveness of leading fine-tuning techniques. Their introduction of a language model boasting 175 billion parameters exemplifies this, as it excels in numerous natural language processing (NLP) tasks and benchmarks, including zero-shot, one-shot, and few-shot scenarios. Furthermore, the model consistently generates high-quality outputs and exhibits robust qualitative performance, even in spontaneously formulated tasks. Despite encountering some limitations and challenges, these findings underscore the potential significance of very large language models in the development of versatile, general-purpose language systems.

Lu et al. [142] introduced Error Analysis Prompting (EAPrompt) to enhance large language models' (LLMs) performance in assessing machine translation quality, achieving notable improvements at both system and segment levels. Their method, EAPrompt, exhibits potential for broader application in language generation tasks beyond translation. Using the WMT22 metrics shared task dataset, they employed a standard meta-evaluation approach, utilizing pairwise accuracy for system-level evaluations and the acc * eq variant for segment-level assessments. The findings demonstrated that EAPrompt significantly enhanced LLM performance at the system level, surpassing other metrics and strategies. Additionally, EAPrompt outperformed GEMBA in most segment-level evaluations and effectively distinguished between major and minor errors, aligning well with the human evaluation framework MQM.

Moreover, Peng et al. [143] examined strategies to enhance ChatGPT's translation capabilities through adjustments in temperature settings and the application of task-specific and domain-specific prompts, leading to improvements especially in complex language pairs. Their study assessed how different temperature settings, the use of specific prompts, and advanced in-context learning methods like few-shot learning and Chain-of-Thought prompting influence translation accuracy. The results showed that optimal temperature settings and focused prompts significantly improve translation performance, but noted that Chain-of-Thought prompting could negatively impact translation quality by inducing word-by-word translation tendencies. Simultaneously, Hendy et al. [144] conducted an assessment of GPT models for machine translation, highlighting their advantages and drawbacks across various levels of language resources. Their primary observations underscore the commendable translation quality of GPT models for languages with abundant resources, their constrained effectiveness for languages with limited resources, the possibility of hybrid methodologies, and the optimistic translation potential of GPT models despite variances in architecture and training data.

Jiao et al. [145] conducted an initial assessment of ChatGPT for machine translation, highlighting its advancements with the GPT-4 engine and positioning it as a proficient trans-

lator. Their methodology involved evaluating ChatGPT's performance in machine translation, with a focus on translation prompts, multilingual translation, and translation robustness. This evaluation encompassed the assessment of prompts, multilingual translation proficiency, and translation resilience across specific test sets. The primary findings revolved around ChatGPT's performance in machine translation, demonstrating competitive outcomes with commercial products for languages with abundant resources, notable enhancement with GPT-4, and reduced errors compared to GPT-3.5.

Pourkamali et al. [140] conducted a thorough investigation into various large language models (LLMs) for machine translation, focusing on their strengths and weaknesses. They highlighted models like PaLM and Perplexity AI for their ability to produce human-like translations and process lengthy texts efficiently, while also discussing the limitations of models such as GPT 3.5. The study emphasized the importance of prompt adjustments and additional data for improving translation quality. Through comprehensive evaluation across different language pairs, the study revealed that LLMs, especially those trained with multilingual data like PaLM, show promise in generating high-quality translations and adapting to various translation nuances. The findings underscored the significance of factors such as training data, prompting methods, and adaptability in influencing LLM performance in machine translation tasks.

Jiang et al. [149] conducted an assessment of ChatGPT and NMT engines in translating Chinese diplomatic texts into English, emphasizing the importance of customized prompts and the superior performance of ChatGPT, particularly when provided with relevant examples or contextual information. Their approach involved evaluating translation quality using both automated metrics and human assessment based on error types and six analytical rubrics. The study addressed specific research questions and extensively analyzed error penalties, severity levels, error categories, and human evaluator ratings across the six rubrics. Results indicated that ChatGPT outperformed NMT systems in human evaluation and semantic-aware automated assessment, with the inclusion of examples or contextual details significantly enhancing its translation accuracy. Conversely, NMT systems exhibited comparable performance, and the correlation between automated metrics and human evaluation was weak and statistically insignificant.

4.6 Summary

In this chapter, we delved into the transformative impact of large language models (LLMs), notably focusing on the advancements brought forth by models such as Transformers, BERT, and GPT within the realm of natural language processing (NLP). These models, built upon

Transformer architectures, have revolutionized NLP tasks by learning intricate linguistic patterns and relationships. Particularly, BERT and GPT represent pioneering models, with BERT excelling in understanding bidirectional context and GPT specializing in generative tasks.

Leveraging LLMs for machine translation (MT), variants like BERT, T5, and ChatGPT demonstrate promising capabilities in bridging language barriers and enhancing translation accuracy. However, challenges persist, including the need for diverse and extensive training data and mitigating biases inherent in these models. Nevertheless, ongoing research endeavors continue to explore advancements and refine techniques in leveraging LLMs for MT, underscoring the dynamic evolution of NLP and MT technologies.

PART II:
DATASET CREATION AND EXPERIMENTS

Chapter 5

Algerian Arabic Corpus by Data Augmentation

5.1 Overview

In this chapter, we delve into the vital significance of corpora and data augmentation techniques in bolstering machine translation (MT) systems. We initiate our discussion by focusing on Algerian Arabic dialectal corpora, elucidating their significance and detailing existing data sources. Subsequently, we delve into the essence of both monolingual and bilingual corpora, elucidating existing data sources in the process. Furthermore, we underscore the critical role of data preprocessing, delving into methodologies aimed at refining corpora to uphold quality and ensure consistency throughout the dataset.

We then delve into monolingual corpora augmentation methods, including copied-corpus augmentation (CC) and back-translation augmentation (BT), which leverage additional data to enhance model performance. Furthermore, we introduce two novel augmentation strategies tailored to Arabic languages: Right Rotation Augmentation (RRA) and Entity Replacement Augmentation (ERA). These approaches aim to address challenges in low-resource language translation by diversifying datasets and incorporating culturally relevant substitutions. Throughout, we underscore the critical role of effective data preprocessing and augmentation in optimizing MT systems for improved translation accuracy and linguistic understanding.

5.2 Algerian Arabic dialect

The Maghrebi dialects, which include Algerian Arabic, are predominantly derived from Standard Arabic, although not exclusively. However, due to practical constraints, several morpho-syntactic rules of Standard Arabic are not consistently adhered to in these Arabic dialects.

This presents a challenge when adapting existing Natural Language Processing (NLP) resources designed for Standard Arabic to handle Arabic dialects [150]. Moreover, numerous studies have highlighted that tools specifically designed for Modern Standard Arabic (MSA) exhibit significantly reduced efficacy when applied to texts written in Arabic dialects, primarily due to the substantial linguistic differences between MSA and these dialects [151]. The Algerian Arabic dialect, also known as Darija, despite its widespread usage, represents a low-resource language with limited parallel data available for machine translation (MT). Similarly, Modern Standard Arabic (MSA), the formal language utilized in media and education in Algeria, also qualifies as a low-resource language for MT [18]. Figure **5.1**[1] visually represents the geographical distribution of various dialectal Arabic varieties across different countries and their respective borders.

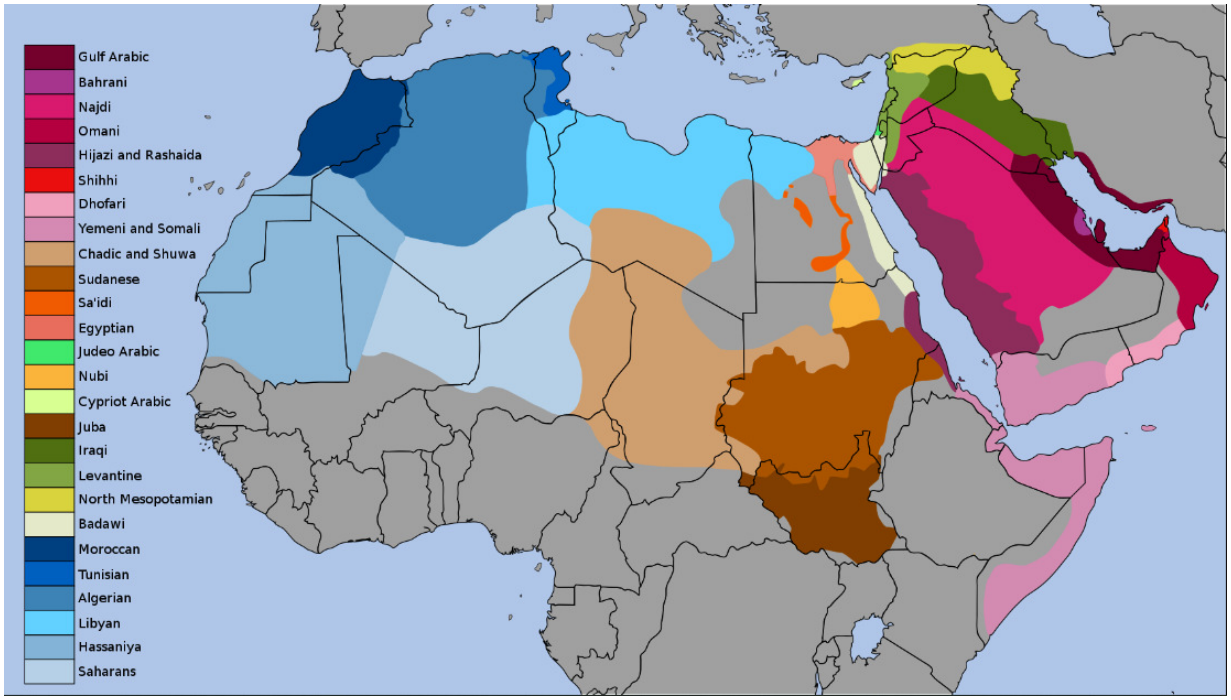

Figure 5.1: Mapping the geographic locations of Arabic Dialect Varieties, including the Algerian Dialect.

The complex morphology and rich inflectional system of both Darija and MSA pose a significant challenge for MT. Few number of research studies have endeavoured to tackle this issue, like the work of [152], who created a collection of parallel data in Algerian Arabic and English for machine translation (MT). Additionally, [153] undertook the development of a diverse parallel Arabic corpus, initially consisting of five Arabic dialects, including two from Algeria, one from Tunisia, and two from the Middle East, alongside the standard Ara-

[1]Map from Wikipedia distributed under a CCBY 3.0 license

Resource Level	Language Pair	Parallel Sentences
High	English–French	280M
Medium	English–Myanmar	0.7M
Low	English–Fon	0.035M

Table 5.1: Examples of language pairs with different levels of resources.

bic. However, this limited availability of parallel data remains a significant hurdle for the development of high-quality MT systems for Algerian Arabic and MSA.

5.3 Bilingual Corpora

Corpus-based approaches to MT, like Statistical Machine Translation (SMT) and Neural Machine Translation (NMT), emerged with the growing availability of parallel corpora or bitexts, which consist of texts translated into different languages. These methods capitalize on the translations created by human translators, utilizing them within machine learning frameworks to construct translation models. High-quality and substantial amounts of parallel data are crucial for training optimal models, as highlighted by [9, 14]. Unlike English, the main issue that faces the Arabic language is the lack of sufficient available datasets; particularly, the parallel datasets; which makes Arabic and dialectal Arabic considered low resource languages (LRLs). The disparities between what can be categorized as high, medium, and low resource language pairs are shown in Table 5.1 [8]. Consequently, a significant challenge in translating low-resource languages lies in obtaining sufficiently large and clean parallel corpora.

DZDA (Algerian Dialectal Arabic) serves as another example of a low-resource language explored in this research. DZDA, a variant of Arabic predominantly spoken in Algeria, is heavily influenced by various linguistic sources, including French, Spanish, and Berber. This linguistic diversity poses challenges for Natural Language Processing (NLP) tasks due to the scarcity of tools and resources tailored to this specific dialect [154]. Moreover, in media content, written texts in DZDA may feature a mix of Arabic script, Latin script, and transliterated words to Latin. This complexity in script and vocabulary adds to the difficulty of processing and analyzing DZDA text, especially in the context of social media content where linguistic variations and informal expressions are prevalent.

The existing corpora (Section 5.5) for DZDA are either limited in size or suffer from poor quality; they are predominantly sourced from online platforms, including crowd translation efforts. Relying on the internet as a corpus repository is driven by the desire to access ex-

tensive data at minimal cost. However, these sources often lack consistency and may contain inaccuracies, posing challenges for machine translation. Moreover, given the absence of standardization in DZDA, variations in language usage and style are common across different sources. As a result, manual or automated methods for data cleaning and alignment are necessary to address these issues and improve the quality of MT outputs.

5.4 Monollingual Corpora

Statistical Machine Translation (SMT) relies on monolingual corpora to develop language models. Although not mandatory for Neural Machine Translation (NMT), monolingual corpora can aid in generating synthetic parallel sentences through techniques like back-translation (BT), where monolingual data in the target language is utilized, and forward-translation (FT), where the monolingual corpus is in the source language.

5.5 Existing Corpora

There are two free datasets available for the task of machine translation between Algerian Arabic dialect and Modern Standard Arabic (MSA): PADIC[2] (Parallel Arabic Dialect Corpus) [153], MADAR[3] (Multi Arabic Dialect Applications and Resources) [152], and ANMaT, our in-house dataset[4]. Table 5.2 summarizes the statistics for each dataset, encompassing metrics such as the quantity of parallel sentences, overall word count, vocabulary size, and average sentence length.

5.6 Data Sources

PADIC encompasses a collection of five dialects: one from Syria, one from Tunisia, one from Palestine, and two originating from Algeria (**Algiers**, **Annaba**), constituting a total of 6,400 parallel sentences for each dialect. The MADAR dataset comprises 25 Arabic dialects, featuring cities such as Beirut, Cairo, Doha, Rabat, Tunis, Aleppo, Alexandria, **Algiers**, Amman, Aswan, Baghdad, Basra, Benghazi, Damascus, Fes, Jeddah, Jerusalem, Khartoum, Mosul, Muscat, Riyadh, Salt, Sanaa, Sfax, and Tripoli, each containing 2,000 sentences. Thus, the total size of MADAR amounts to 50,000 sentences.

[2]PADIC dataset is available at: https://sourceforge.net/projects/padic/
[3]MADAR dataset is available at: https://github.com/farahshamout/madar-dataset
[4]an internally created dataset is available at: https://github.com/bbaligh/DZDA-MSA/

ANMaT, the in-house dataset, was created for the purpose of enhancing MT efforts, was meticulously assembled by two expert native speakers fluent in both the Algerian Arabic Dialect (DZDA) and Modern Standard Arabic (MSA). Using a specialized web scraping tool, a rich array of bilingual sentence pairs was compiled from diverse social media sources, such as Twitter, YouTube, and Facebook. Adhering to a strict one-to-one correspondence, each sentence from one language was carefully translated to the other, with thorough preprocessing to guarantee the uniformity and quality of the data. The engagement of two native speakers contributed to the dataset's linguistic accuracy, enriching it with cultural nuances and idiomatic phrases unique to the Algerian Arabic Dialect and Modern Standard Arabic. To protect user privacy, anonymization measures were taken, ensuring the in-house dataset's data remained confidential and secure throughout the translation effort. The collection consists of around 1,800 sentence pairs, featuring parallel texts in DZDA and MSA.

Dataset	Language	# Tokens	Vocabulary	# Sentences	Average length
PADIC	MSA	87,680	8,374	6,412	13.65
	DZDA	78,614	7,613		12.26
MADAR	MSA	15,929	4,408	2,000	10.23
	DZDA	13,198	4,180		8.56
ANMaT	MSA	17,594	2,632	1,800	12.09
	DZDA	17,877	3,200		12.23
Consolidated Dataset	MSA	132,512	11,492	10,212	14.04
	DZDA	119,664	11,927		12.71

Table 5.2: Statistic of the MSA↔DZDA corpora

5.7 Preprocessing

The dataset pre-processing steps outlined in the algorithm 1 encompass a series of cleaning and selecting procedures to prepare the data for further analysis. Initially, the cleaning phase involves the removal of non-alphanumeric characters from the dataset to ensure data cleanliness. Subsequently, sentences exhibiting a length ratio greater than 3 or less than 0.3 are eliminated, as they are considered outliers in terms of length discrepancy. Additionally, sentences with a lexical overlap of less than 0.5 are removed to enhance the dataset's lexical consistency.

Algorithm 1: Dataset pre-processing steps

1: /* Dataset cleaning */
2: Remove non-alphanumeric characters
3: Remove sentences with a length ratio > 3 or < 0.3
4: Remove sentences with a lexical overlap < 0.5

5.8 Data Augmentation

NMT is extremely data-hungry [155, 102, 1] and the presence of abundant, high-quality parallel data is essential for achieving the best outcomes. So, when working with LRLs, it is critical to employ approaches that increase corpora size in order to attain better MT quality. There are numerous approaches to increasing the size of the corpora and improving model performance such as: backward translation (henceforth BT) [17, 156] which is an approach where a backward model is used to generate hypotheses of the source language in order to increase the amount of data available to translation systems, forward translation (henceforth FT) [157] which works in the opposite direction, employing a forward model to predict translations in the target language - a process known as Self-learning., and the copied-corpus approach, in which target sentences are copied on the source side [158] (called Mix-source) or, inversely, by making a copy of source sentences into the target side [159]. Recent research has shown that different approaches, including back-translation, sub-word units, and adapting NMT systems to low-resource settings [16], can improve NMT performance with low-resource scenarios.

5.8.1 Monolingual corpora Augmentation

5.8.1.1 Copied-corpus Augmentation

This method enhances training datasets by copying existing sentences from the target language to the source language (termed Mix-source) or the other way around. Originally inspired by self-teaching strategies in MT, the approach was first introduced by [158], who proposed the Mix-source method of duplicating target language sentences on the source side (see Figure 5.2). Following this, [159] explored the inverse, where source language sentences are replicated on the target side. This straightforward method boosts the available training material by reusing the corpus in new ways, broadening the model's exposure to different translation possibilities and linguistic patterns. Such exposure can significantly enrich the model's performance across various MT tasks.

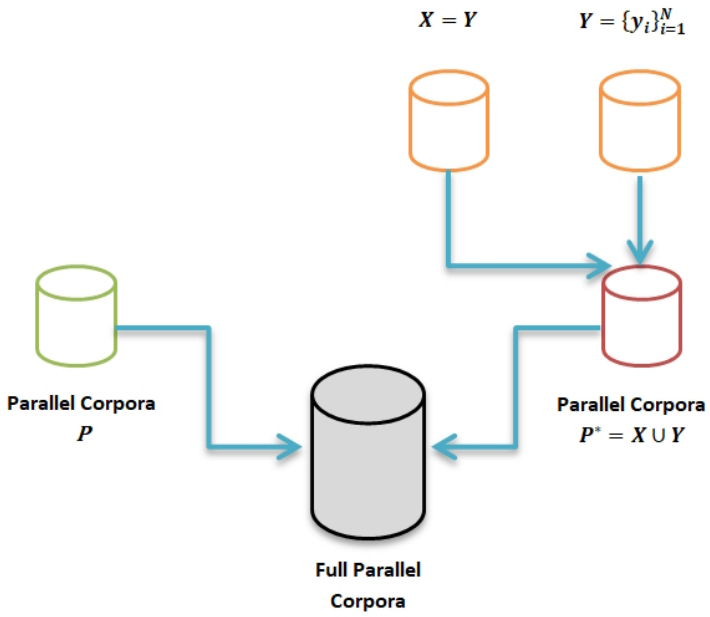

Figure 5.2: The copied-corpus augmentation approach

The advantages of copied corpus augmentation extend beyond simply enlarging the dataset. It notably addresses the challenge of limited data resources, a frequent obstacle in MT, especially with under-represented languages. By multiplying the available examples, it enables the model to better generalize from its training, enhancing its capability to deal with novel vocabulary or sentence structures. This method is especially relevant for neural machine translation systems that thrive on recognizing data patterns. Replicating sentences across language boundaries may reinforce pattern recognition, thereby increasing translation precision.

Nevertheless, this technique is not without its potential downsides. The act of copying sentences might not always yield semantically rich training examples, requiring careful strategy selection to avoid injecting noise into the dataset. Furthermore, its efficacy can fluctuate based on the specific language combination and the architecture of the MT system being used. Despite these considerations, copied corpus augmentation stands as a notably straightforward yet potent strategy for enhancing data volume in machine translation. Its capacity to mitigate data scarcity issues and boost model performance renders it an invaluable asset for MT researchers and developers.

5.8.1.2 Back Translation Augmentation

Enhancing Neural Machine Translation (NMT) quality can be achieved by leveraging extra monolingical resources to generate synthetic training data. Typically, monolingual data on

the source side is translated into the target language (FT), while target-side monolingual data undergoes back BT. This synthetic data is then incorporated with the initial bilingual corpus.

BT involves translating sentences from a target language back into the source language using an existing NMT model. These back-translated sentences are then paired with their original target language sentences, effectively creating new, synthetic source-target sentence pairs (see Figure 5.3). This method significantly enriches the training dataset, especially with examples that might not be present in the original parallel corpus. By leveraging monolingual data in the target language, back translation helps in bridging the gap in data scarcity and improves the NMT model's ability to understand and translate nuanced and complex sentence structures. This technique has been widely acknowledged for its capacity to boost the quality of machine translations, making it a favored choice among researchers and practitioners aiming to enhance the performance of MT systems, particularly in scenarios involving low-resource languages. It's commonly acknowledged that BT significantly enhances Neural Machine Translation more effectively than FT [160].

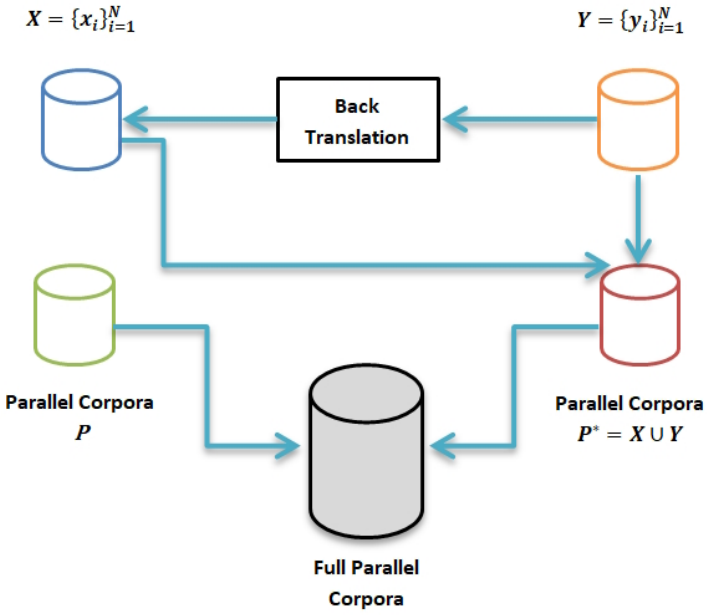

Figure 5.3: The back-translation augmentation approach

5.8.2 Parallel Corpora Augmentaiton

5.8.2.1 Right Rotation Augmentation

After exploring monolingual augmentation strategies such as copied-corpus augmentation and back-translation, we now introduce a novel method tailored to the unique syntactic properties

of Arabic and implemented on bilingual corpora, dubbed "Right Rotation Augmentation". This technique leverages the flexible word order in Arabic sentence construction, where, due to its non-configurational nature, elements within a sentence can often be rearranged without altering the sentence's meaning.

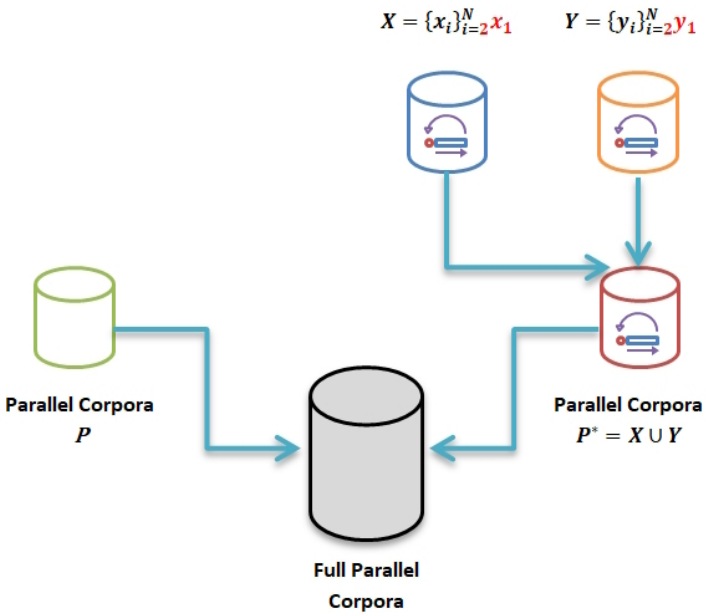

Figure 5.4: The Right-rotation augmentation approach

Right Rotation Augmentation systematically rotates the sentence structure to the right, creating multiple valid syntactic variations of the same sentence. For instance, a sentence beginning with a verb can be transformed to start with its object or subject without losing coherence (see Figure 5.4). This method not only enriches the dataset with diverse syntactic representations but also helps machine translation models better grasp the variability and richness of Arabic syntax. By incorporating such rotated sentences into the training data, models can learn to recognize and translate a wider array of sentence constructions, potentially improving translation accuracy and robustness, especially for languages with free or flexible word order like Arabic.

The Right Rotation Augmentation (RRA) algorithm (see Algorithm 2) takes a sentence pair as input and applies right rotation to both the source and target sentences, generating four augmented sentence pairs with varied word orders. This approach introduces diversity into the training data, potentially enhancing the robustness and performance of machine translation models. The new size of the resulting augmented dataset using RRA is four times the original size of the dataset.

Algorithm 2: Right Rotation Augmentation (RRA)

Input: A sentence pair S_1-T_1 (where S_1 represents the source sentence and T_1 denotes the target sentence)

Output: Four augmented sentence pairs: S_1-T_1, S'_1-T_1, S_1-T'_1, and S'_1-T'_1

/* Apply Right Rotation to S_1 ;

$S'_1 \leftarrow$ Rotate the source sentence S_1 by moving the first word to the end;

/* Apply Right Rotation to T_1 ;

$T'_1 \leftarrow$ Rotate the target sentence T_1 by moving the first word to the end;

return *Original sentence pair* S_1-T_1*, Augmented pair* S'_1-T_1*, Augmented pair* S_1-T'_1*, Augmented pair* S'_1-T'_1

5.8.2.2 Entity Replacement Augmentation

A novel augmentation method, termed Entity Replacement Augmentation (ERA) or Lexicon-based Entity Substitution Augmentation, has been created specifically for low-resource languages, including Algerian Dialectal Arabic (DZDA) and Modern Standard Arabic (MSA) in our case. This method involves substituting entities, such as person or location names, using a predefined lexicon. By replacing these entities with alternate names from the lexicon, fresh sentences can be generated while preserving the original sentences' structure and context. This method aims to enrich the dataset by introducing variations in the mentioned names, potentially enhancing the model's capacity to handle diverse linguistic scenarios. Moreover, by integrating culturally relevant names into the augmented dataset, the approach seeks to enhance the model's grasp of context-specific language usage, thereby contributing to more precise and culturally attuned translations.

To execute this task, first, we need to compile a lexicon containing alternative names for entities commonly found in the text, such as names of people, places, organizations, and other relevant entities. This lexicon can be sourced from various resources, including dictionaries, databases, or even generated using statistical methods based on the existing dataset. Once the lexicon is prepared, we identify the entities within the sentences that we intend to augment. For each identified entity, we randomly select a replacement name from the lexicon. Finally, we substitute the original entity with the chosen replacement name, generating new sentences with the altered entities. This process is repeated iteratively for multiple sentences in the dataset, resulting in an augmented dataset with variations in entity names.

Algorithm 3: Entity Replacement Augmentation

Input: Sentence pairs $S1 - T1$ (S1: source sentence, T1: target sentence), Lexicon containing alternative names for entities

Output: Augmented sentence pairs

Input : Sentence pairs $S1 - T1$ (S1: source sentence, T1: target sentence), Lexicon containing alternative names for entities

Output: Augmented sentence pairs

```
for each sentence pair (S1, T1) in the dataset do
    Identify entities in S1 and T1;
    for each identified entity in S1 do
        Select a random replacement name from the lexicon;
        Replace the entity in S1 with the chosen replacement name;
    end for
    for each identified entity in T1 do
        Select a random replacement name from the lexicon;
        Replace the entity in T1 with the chosen replacement name;
    end for
    Append the original sentence pair (S1, T1) and the modified pair (S'1, T'1) to
     the augmented dataset;
end for
```

5.9 Summary

In conclusion, this chapter has explored various strategies aimed at optimizing machine translation (MT) through the utilization of different types of corpora and data augmentation techniques. We began by highlighting the significance of monolingual and bilingual corpora as essential resources for MT tasks, emphasizing their role in training robust models. Subsequently, we delved into the preprocessing steps necessary for refining corpora to ensure data quality and consistency, emphasizing the importance of thorough cleaning and normalization procedures.

Furthermore, we investigated monolingual corpora augmentation methods, focusing on copied-corpus (CC) and back-translation (BT) techniques, which leverage additional monolingual data to augment training sets and improve model performance. Additionally, we introduced two novel augmentation approaches tailored to Arabic languages: Right Rotation Augmentation (RRA) and Entity Replacement Augmentation (ERA). RRA involves rotating sentence structures to diversify datasets, while ERA substitutes entities like names with

culturally relevant alternatives, enriching training data. These strategies aim to address challenges posed by low-resource languages and enhance translation accuracy. Throughout the chapter, we emphasized the significance of effective data preprocessing and augmentation in optimizing MT systems for improved linguistic understanding and translation capabilities.

Chapter 6

Enhancing NMT Using Data Augmentation Techniques

6.1 Overview

In this chapter, we delve into the realm of enhancing Neural Machine Translation (NMT) systems using data augmentation techniques. NMT has revolutionized the field of machine translation, but its performance can be hindered, especially in low-resource language pairs. To address this challenge, we explore various data augmentation strategies aimed at enriching the training data and improving translation quality.

In the Section 6.2, we begin by presenting the system architecture employed for our experiments. This section outlines the underlying framework used for training and evaluating the NMT models, providing insights into the components and configurations essential for the augmentation process.

Next, we establish a baseline NMT system for the target language pair (Section 6.3). This baseline serves as a reference point against which the performance of augmented models is compared. We describe the training methodology and parameters used to build the baseline system, laying the foundation for subsequent experiments.

With the baseline system established, we proceed to experiment with various data augmentation techniques. We explore methods such as Back Translation, Copied Corpus, and novel approaches like Right Rotation Augmentation, aiming to augment the training data and enhance the robustness of the NMT models. This section (Section 6.5) outlines the experimental setup, including the selection of augmentation techniques, dataset preparation, and evaluation metrics.

Finally, in the Section 6.6, we present the results of our experiments and engage in a comprehensive discussion. We analyze the performance of augmented NMT models compared

to the baseline, examining metrics such as BLEU scores and qualitative aspects of translation quality. Through critical evaluation and interpretation of the results, we aim to gain insights into the effectiveness of different augmentation techniques and their implications for improving NMT in low-resource language scenarios.

By exploring these key aspects, this chapter sheds light on the potential of data augmentation techniques to enhance NMT systems and lays the groundwork for further advancements in machine translation research.

6.2 System Architecture

In the current study, the foundational model employed is predicated on the well-established sequence-to-sequence (seq2seq) framework, as detailed in seminal works by [155, 161]. This framework is inherently designed to convert sequences from an input domain to a corresponding sequence in a target domain, proving pivotal for tasks such as machine translation. Central to the effectiveness of this model is the integration of an attention mechanism, initially conceptualized by [102], which enhances the model's ability to selectively concentrate on specific segments of the input sequence that are most salient for generating the subsequent element in the output sequence. This selective attention is facilitated through a dynamic alignment score that guides the model's focus across different parts of the input sequence during the decoding process (see Figure 7.1).

The architectural design of this model incorporates an encoder and a decoder, each composed of 300 gated recurrent units (GRUs). GRUs are a type of recurrent neural network (RNN) architecture known for their efficiency in capturing temporal dependencies and managing longer sequences without succumbing extensively to issues like vanishing gradients, which are more prevalent in traditional RNNs. The configuration of these GRUs within the encoder-decoder structure allows for a robust handling of sequence data, adapting dynamically to the length and complexity of the input and target sequences.

For the training of the neural machine translation (NMT) model, specific parameters were adhered to, as detailed in the referenced Table 6.2. The training process followed a methodical approach outlined in Algorithm 4, ensuring a systematic and reproducible procedure for model optimization. The choice of these parameters was strategically made to optimize the trade-off between computational demand and the performance of the model, aiming to achieve high translation accuracy while maintaining manageable computational loads.

The experimental setup for evaluating the model's performance utilized the TensorFlow deep learning framework, a popular choice for its flexible and comprehensive tools that fa-

cilitate both the design and training of complex neural network architectures. The computational experiments were executed on Google Colab, leveraging the computational power of a Tesla T4 GPU with 16GB of RAM. This setup provided a controlled environment for rigorous testing and evaluation of the model, ensuring that the results are both robust and reliable, reflective of both the potential and limitations of the employed architectural and operational configurations.

Algorithm 4: Seq2Seq NMT Model training

```
EarlyStop ← 5 ;                    /* Early stopping mechanism = 5 epochs */
NbIdenticalLoss ← 1;
Iteration ← 0;
while (Iteration ≤ Epochs) and (NbIdenticalLoss ≤ Early_Stop) do
    Train Model;
    Evaluate Loss;
    if LossValue = LastLossValue then
        NbIdenticalLoss + +;
    else
        NbIdenticalLoss ← 1;
        LastLossValue ← LossValue;
    end if
    Iteration + +;
end while
```

6.3 Baseline System

The cornerstone of every NMT model depends on the quality and quantity of the training data it utilizes. Algerian dialect resources are extremely limited and are only available in MADAR, PADIC, Tatoeba, and an in-house dataset corpora[18]. MADAR (Multi Arabic Dialect Applications and Resources) encloses Arabic dialects from 25 Arab cities (including Algiers), each with 2000 sentences in its version CORPUS-25 [152]. PADIC (A Parallel Arabic Dialect Corpus) includes five dialects (two from Algeria: Algiers and Annaba), each with 6,400 parallel sentences [153]. Tatoeba[1] is a collection of 421 languages' sentences and translations, including 53,786 Arabic MSA and 2,357 Algerian dialect sentences. The in-house dataset[2], which was generated by the study's authors, has a total of about 1,800 sentence pairings and includes parallel sentences in MSA and Algerian dialects. Table 6.1

[1]https://tatoeba.org (Retrieved 05/18/2023)
[2]https://github.com/bbaligh

Table 6.1: Statistics of the utilized MSA-DZDA datasets

Dataset	Sentences	DZDA			MSA		
		Tokens	Vocab.	Average length	Tokens	Vocab.	Average length
Baseline	10,212	119,664	11,926	12.71	132,512	11,491	14.04
Copied Corpus	63,998	652,699	41,494	11.77	665,547	37,912	11.99
BT. Corpus	63,998	601,017	39,347	10.86	665,547	37,912	11.99
RR. Corpus	37,169	435,604	11,926	12.59	480,588	11,491	13.87

provides some statistics on the parallel corpus used as baseline for training (i.e. without augmented/synthetic data) The copied corpus is constructed by using the baseline dataset and the monolingual MSA dataset from the Tatoeba corpus for the Copied-Corpus data augmentation.

The objective of the Right-Rotation technique was to maintain the integrity of the original sentence pairs while introducing variations. The new dataset comprised four types of sentence pairs. The first type retained the original source and target sentences without alterations. In the second type, the source sentence remained unchanged, while the target sentence was created by appending the initial word of the original target sentence to the end, provided it exceeded three characters in length. Similarly, the third type preserved the original target sentence while rotating the source sentence according to the same rotation condition. Lastly, the fourth type combined the rotated source and target sentences. Rotation was omitted if the length of the initial word was less than three characters. This augmentation method aimed to augment the dataset's diversity and variability, potentially improving the performance of language models trained on the augmented data.

6.4 Segmentation

6.4.1 Word Segmentation

Word tokenization is a fundamental text segmentation strategy employed in neural machine translation (NMT) tasks. This method splits source and target language sentences into individual words based on spaces or punctuation marks [4]. Its simplicity offers several advantages: it is computationally efficient, making it suitable for resource-limited environments, and the resulting tokens directly correspond to words, enhancing interpretability and simpli-

fying the analysis of the model's processing. However, word tokenization also has limitations. One significant challenge is its vulnerability to out-of-vocabulary (OOV) words. The model may struggle to translate words it has not encountered during training, leading to potential errors in the translated output. Additionally, word tokenization can be language-dependent. Languages with ambiguous word boundaries or complex morphology, such as agglutinative languages where words are formed by adding suffixes, present challenges for this approach. Various tools are available for text tokenization, each serving different languages and purposes:

- Moses Tokenizer [162]: It is included with the Moses toolkit. It separates punctuation from words while preserving special tokens such as URLs or dates and normalizes characters (e.g., different Unicode variants of quotes). It is applicable to any language. Figure 6.1.a presents an example of tokenizing a sentence using the *Moses Tokenizer*.

- OpenNMT Tokenizer [163]: It is included with the OpenNMT toolkit, which normalizes characters (e.g., different Unicode variants of quotes) and separates punctuation from words. It is versatile and can be used with any language. Figure 6.1.b shows an example of tokenizing a sentence using the *OpenNMT Tokenizer*.

- SentencePiece[3]: An unsupervised text tokenizer and detokenizer designed primarily for neural network-based text generation systems, where the vocabulary size is fixed before training. It is applicable to any language, although separate models need to be trained for each language. Figure 6.1.c shows an example of tokenizing a sentence using *SentencePiece*.

6.4.2 Subword Segmentation

Subword tokenization is the standard approach for tokenization in neural language models and machine translation systems [164]. This technique is a prevalent method used to facilitate open-vocabulary translation by encoding rare words with sequences of subword units. This approach enables NMT models to translate or generate previously unseen words during inference while effectively reducing the vocabulary size of the entire training dataset. In this study, subword units were learned by applying Byte Pair Encoding (BPE) [17] on the combined source and target corpora (i.e., joint BPE segmentation). The original vocabulary sizes of the MSA and DZDA baseline training data were 11,491 and 11,926 tokens, respectively. After applying BPE, the common vocabulary size was empirically set to 15,000 pieces.

[3]https://github.com/google/sentencepiece

Original: *In a browser window (Internet Explorer or Firefox) browse to www.dellconnect.com.*
Segmented: *In a browser window (Internet Explorer or Firefox) browse to www.dellconnect.com.*

(a) Example of a sentence tokenized using *Moses tokenizer*. The tokenization has split the punctuation, without modifying the url.

Original: *In a browser window (Internet Explorer or Firefox) browse to www.dellconnect.com.*
Segmented: *In a browser window (Internet Explorer or Firefox) browse to www . dellconnect . com .*

(b) Example of a sentence tokenized using *OpenNMT tokenizer*. The tokenization has split punctuation and transformed the url into several words.

Original: *In a browser window (Internet Explorer or Firefox) browse to www.dellconnect.com.*
Segmented: *_In _a _browser _window _(Internet _Explorer _or _Firefox) _browse _to _www . dell connect . com .*

(c) Example of a sentence tokenized using *SentencePiece*. _ indicates the start of a word in the original sentence. The tokenization has split punctuation and transformed the url into several words.

Figure 6.1: Examples of segmenting sentences with each word segmenter, adapted from [4].

6.5 Experiments and Evaluation

The experimental setup for evaluating the model's performance utilized the TensorFlow deep learning framework, a popular choice for its flexible and comprehensive tools that facilitate both the design and training of complex neural network architectures. The computational experiments were executed on Google Colab, leveraging the computational power of a Tesla T4 GPU with 16GB of RAM. This setup provided a controlled environment for rigorous testing and evaluation of the model, ensuring that the results are both robust and reliable, reflective of both the potential and limitations of the employed architectural and operational configurations. For the training of the neural machine translation (NMT) model, specific parameters were adhered to, as detailed in the referenced Table 6.2.

Table 6.2: Values of the Seq2Seq models hyperparamters

Parameter	GRU units	Optimizer	Learning rate	Batch size	Epochs	Dropout
Value	300	Adam	0.001	24	50 (max)	0.1

We assess the translation quality by comparing the predictions with the ground truth using BLEU [49] Case-insensitive and detokenized BLEU scores are computed using SacreBLEU [165].

6.6 Results and Discussions

The results of our experiments demonstrate notable improvements in translation quality when employing data augmentation techniques for both the DZDA-to-MSA and MSA-to-DZDA translation tasks. Comparing the baseline performance with that of the augmented corpora reveals substantial enhancements across the board. The BLEU scores of the NMT system trained on distinct corpora are reported in Table 6.3

For the DZDA→MSA translation task, the baseline BLEU scores are 19.1 for word tokenization and 20.4 for subword segmentation. Introducing the Copied Corpus augmentation technique results in a substantial improvement, with BLEU scores increasing to 31.1 for word tokenization and 33.2 for subword segmentation. Similarly, employing the BT Corpus augmentation yields notable enhancements, achieving BLEU scores of 30.0 for word segmentation and 32.1 for subword segmentation. Although slightly lower, the RR Corpus augmentation still demonstrates significant progress, with BLEU scores of 27.1 for word tokenization and 28.9 for subword segmentation.

Table 6.3: MSA-DZDA translation performance using BLEU score

Dataset	Segmentation Unit	DZDA→MSA	MSA→DZDA
Baseline	word	19.1	17.7
	subword	20.4	18.8
CC-Corpus	word	31.1	30.1
	subword	**33.2**	**31.9**
BT-Corpus	word	30.0	28.8
	subword	32.1	30.5
RR-Corpus	word	27.1	26.5
	subword	28.9	28.1
CC-Corpus+BT-Corpus	word	31.2	20.8
	subword	33.5	22.1
CC-Corpus+RR-Corpus	word	31.9	31.6
	subword	**33.9**	**33.4**
BT-Corpus+RR-Corpus	word	28.8	21.3
	subword	28.9	28.1

For the MSA→DZDA translation task, we observe a similar pattern. The baseline BLEU scores are 17.7 for word segmentation and 18.8 for subword segmentation. Implementing the Copied Corpus augmentation technique leads to a significant increase, resulting in BLEU scores of 30.1 for word tokenization and 31.9 for subword tokenization. Likewise, the BT Corpus augmentation shows substantial enhancement, yielding BLEU scores of 28.8 for word segmentation and 30.5 for subword segmentation. The RR Corpus augmentation, though slightly lower, still shows noteworthy improvements, with BLEU scores of 26.5 for word tokenization and 28.1 for subword segmentation.

The results in Table 6.3 further underscore the potential of combined data augmentation techniques in enhancing translation quality. For the DZDA→MSA translation task, the combination of Copied Corpus and Right-Rotation (CC+RR) achieved the highest BLEU scores, with subword units reaching 33.9 and word units at 31.9. Similarly, for the MSA→DZDA translation direction, the CC+RR approach also outperformed other combinations, yielding BLEU scores of 33.4 with subword units and 31.6 with word units. These results indicate that the integration of multiple augmentation strategies, particularly the CC+RR combination, can significantly enhance translation performance, especially when subword segmentation is employed. This underscores the importance of selecting effective augmentation methods and optimizing the unit of segmentation to maximize translation quality.

The "Right-Rotation Augmentation" (RRA) strategy presents both advantages and drawbacks, delineated as follows:

Advantages:

- Enhanced Data Diversity: RRA generates novel sentences by rotating existing ones, thus enriching the corpus with a broader array of sentence structures and word arrangements. This heightened diversity can enhance the adaptability and generalization capabilities of the neural machine translation (NMT) model.

- Preservation of Sentence Structure: RRA maintains the fundamental structure of sentences while altering word order, ensuring that the augmented sentences uphold the core elements of Arabic language construction. This preserves grammatical coherence and integrity within the augmented corpus.

- Simplicity and Efficiency: RRA is a straightforward and efficient data augmentation method that can be readily implemented. It does not necessitate additional external resources or intricate preprocessing procedures, rendering it accessible and time-effective for augmenting modest-sized MT corpora.

Drawbacks:

- Limited Semantic Variation: While RRA diversifies sentence structures, it may not introduce significant semantic deviations in the augmented sentences. The meaning and content of sentences might remain relatively consistent, potentially constraining the NMT model's capacity to discern and handle nuanced semantic nuances.

- Potential for Unnatural Sentences: Depending on the extent of rotation and sentence complexity, RRA may generate sentences that appear unnatural or less commonplace in everyday language usage. This could affect the quality and fluency of translations produced by the NMT model.

- Data Sparsity: RRA does not directly address data scarcity concerns. Despite generating new sentences, the augmented corpus size remains constrained by the initial corpus size. In cases where the initial corpus is limited, the augmented corpus may not substantially increase training data.

It's essential to recognize that the efficacy of the RRA approach may vary based on specific implementation intricacies, the quality of the initial corpus, and the linguistic characteristics of the language under augmentation. Evaluating the impact of RRA through empirical assessments and comparative analyses with other data augmentation techniques is advisable. These results suggest promising avenues for addressing the challenges posed by low-resource languages in machine translation tasks. Additionally, exploring further augmentation techniques and assessing their robustness across diverse language pairs and domains could significantly advance the field of low-resource machine translation. This would not only enhance the effectiveness of translation systems for dialectal Arabic and Modern Standard Arabic but also provide insights into improving machine translation for other low-resource language pairs.

6.7 Summary

In this chapter, we explored methods to enhance Neural Machine Translation (NMT) systems through data augmentation techniques. We began by outlining the system architecture used for our experiments, establishing a baseline NMT system for the target language pair. Experimenting with various augmentation strategies, including Back Translation, Copied Corpus, and novel approaches like Right Rotation Augmentation, we aimed to enrich the training data and improve translation quality. Through rigorous experimentation and evaluation, we compared the performance of augmented NMT models against the baseline. Results revealed significant improvements in translation quality, as indicated by metrics such as BLEU scores.

Our analysis highlighted the effectiveness of different augmentation techniques in enhancing NMT systems, particularly in low-resource language scenarios. By providing insights into the benefits and challenges of data augmentation for NMT, this chapter underscores its potential to advance machine translation research. As we conclude our exploration, it becomes evident that augmenting training data can substantially enhance the robustness and performance of NMT models, paving the way for more effective cross-lingual communication and information access in diverse linguistic contexts.

Chapter 7

Seq2Seq Neural vs GPT-based Machine Translation

7.1 Overview

In this chapter, we embark on an in-depth examination of the distinctions between Seq2Seq Neural Machine Translation (NMT) models and GPT-based models, two prominent approaches in the field of machine translation. We begin by elucidating the intricate architectural nuances inherent in each approach. The Seq2Seq models, renowned for their utilization of encoder-decoder frameworks with attention mechanisms, are meticulously dissected to uncover the intricacies of their operations. Conversely, GPT-based models, characterized by their pre-trained transformer architecture and autoregressive generation capabilities, are scrutinized to unveil the underlying mechanisms driving their performance.

Building upon this foundation, we introduce the baseline system that forms the backbone of our comparative analysis. We provide comprehensive insights into its construction, detailing the integration of three diverse datasets tailored for the Algerian Arabic dialect (DZDA) to Modern Standard Arabic (MSA) translation task.

Our experimental framework encompasses a diverse array of scenarios, including zero-shot and few-shot prompting, aimed at exploring the efficacy of both Seq2Seq and GPT-based models under varying conditions. To facilitate rigorous evaluation, we employ a suite of robust metrics, including COMET, BLEU, and ChrF scores, to gauge translation quality with precision.

Upon conducting our experiments, we meticulously analyze the results obtained from translating both MSA to DZA and DZA to MSA. Through this analysis, we uncover nuanced insights into the strengths and limitations of each model type. Furthermore, we complement our quantitative analysis with a qualitative human analysis component, which offers invalu-

able qualitative perspectives on the fluency, accuracy, and naturalness of the translation outputs generated by Seq2Seq and GPT-based models.

By delving into the intricacies of these approaches, our chapter aims to provide a comprehensive understanding of the dynamics between Seq2Seq and GPT-based models in the realm of machine translation, offering valuable insights for future research and development endeavours.

7.2 System Architecture

In this section, we delve into the system architecture employed for our MT experiments, focusing on two key methodologies: the Seq2Seq Neural Machine Translation (NMT) model and the GPT-based model, specifically based on ChatGPT prompting. These models represent distinct approaches to addressing the task of MT, each with its unique architectural intricacies and computational strategies. We provide an in-depth exploration of the underlying mechanisms of both approaches, elucidating their strengths and weaknesses in the context of translating between languages. By dissecting the system architectures of the Seq2Seq NMT model and the GPT-based model, we lay the groundwork for a comprehensive understanding of their functionalities and methodologies, facilitating further analysis and evaluation in subsequent sections.

7.2.1 Seq2Seq NMT Model

The model used in this study is based on the sequence-to-sequence framework, which was introduced by [155, 161]. It consists of an encoder-decoder network that uses gated recurrent units (GRU) cells. (see Figure 7.1). The encoder is responsible for learning the input sentence's embeddings and producing a context vector that encapsulates the sentence's essence during the training process. The decoder then generates the target sentence based on this vector, starting with the <**start**> symbol as input.

The process involves providing the previous predicted output (y_{t-1}) and hidden state (d_{t-1}) as input to the decoder for each time step (t), which generates the current output (y_t). This is used to calculate loss and apply gradients before back-propagation. Teacher forcing is used for the training process, while the inference process uses prior predictions, hidden state, and encoder output as input. The model stops when it predicts the <**end**> token, indicating the sentence's end.

The model is improved by adding attention [102], which assigns weights to each word in the source sentence during target word generation. The encoder's hidden states are summed

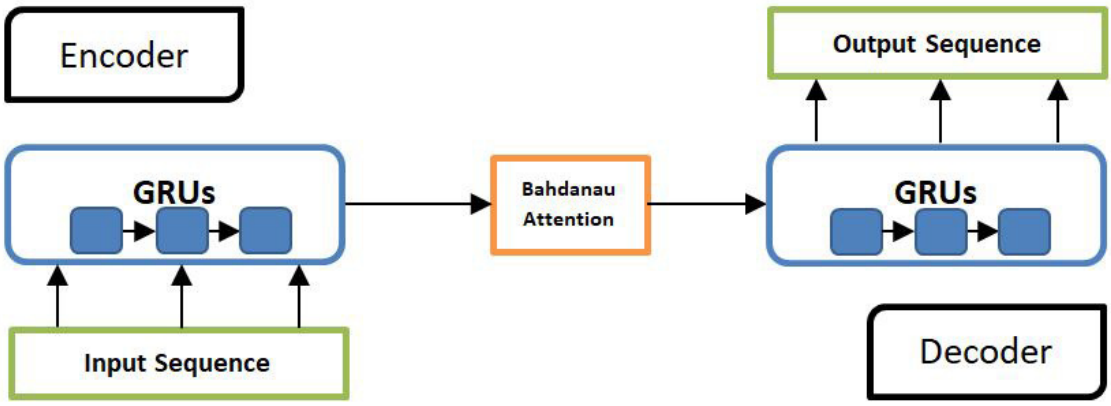

Figure 7.1: The Seq2Seq NMT model

to create the context vector, and attention weights are derived using an alignment function and the softmax function, ranking the hidden states by importance for generating the target word at time (t).

The training of the neural machine translation (NMT) model adhered to specific parameters as detailed in Table 6.2. The process followed a structured approach, as outlined in Algorithm 4, ensuring a systematic and reproducible method for optimizing the model.

7.2.2 GPT based Model

ChatGPT is an advanced language model that originates from OpenAI, built on the enhancements of the GPT-3.5 architecture. It underwent training on an extensive collection of text and code, showcasing robust capabilities across diverse tasks related to natural language understanding. The model has several possible uses, such as text summarization, dialogue systems, creative writing, educational purposes, and research [166]. Additionally, beyond the aforementioned uses, ChatGPT has demonstrated favorable outcomes in machine translation, indicating its effectiveness in grasping the intended significance of sentences.

Furthermore, in addition to its proficiency in prompt engineering, which involves tailoring prompts to guide language models, ChatGPT has also proven to be adept in the realm of machine translation. Leveraging prompt engineering techniques with ChatGPT not only enhances the quality, coherence, and relevance of the generated text but also contributes to its effectiveness in capturing the intended meaning of sentences. This synergy between prompt engineering and ChatGPT's capabilities underscores its versatility in various natural language processing tasks.

In exploring the realm of prompting techniques for language models, various approaches such as zero-shot, one-shot, and few-shot prompting have emerged as valuable methodologies.

Table 7.1: Templates of Zero-shot, One-shot, and Few-shot prompts

Zero-shot	`Translate from Algerian Arabic Dialect DZDA to Arabic MSA:` `[source] =>`

One-shot	`Translate from Algerian Arabic Dialect DZDA to Arabic MSA:`
	`[source] => [target]`
	`[source] =>`

Few-shot	`Translate from Algerian Arabic Dialect DZDA to Arabic MSA:`
	`[source 1] => [target 1]`
	`[source 2] => [target 2]`
	`...`
	`[source n] => [target n]`
	`[source] =>`

7.2.2.1 Zero-Shot Prompting

Zero-shot prompting enables models to make predictions for novel data without the need for additional training, a departure from conventional approaches. In the realm of prompt engineering, it plays a pivotal role in generating natural language text seamlessly, circumventing the necessity for explicit programming. This functionality serves to bolster dynamic text generation models, empowering them to identify and categorize objects autonomously, even in the absence of prior exposure to specific instances. In the context of zero-shot translation, a simple articulation of the translation task in natural language suffices, as illustrated in Table 7.1. This streamlined process underscores the versatility and adaptability of zero-shot techniques in leveraging language models for diverse tasks, facilitating efficient and intuitive interactions with the underlying model architecture.

7.2.2.2 One-Shot Prompting

One-shot prompting involves generating text with minimal input, often a single example. When integrated with dialogue management and context modeling techniques, it significantly improves the performance of text generation systems. Within the realm of prompt engineering, one-shot learning stands out for its ability to yield consistent outputs despite being trained on limited input data.

In the one-shot setup, we introduce a solitary instance of translation either from DZDA to MSA or vice versa (as outlined in Table 7.1), which has been curated by a human translator. This approach enables the model to glean valuable insights from a single example, showcasing

its adaptability and efficiency in producing accurate translations even with sparse training data.

7.2.2.3 Few-Shot Prompting

Few-shot prompting leverages a small dataset of examples to facilitate rapid adaptation of the model to novel instances, a strategy particularly valuable in prompt engineering for generating natural language text with constrained input. Despite its reliance on minimal data, few-shot prompting empowers the development of versatile and adaptive text generation models capable of producing contextually relevant outputs across diverse scenarios. Integration of advanced methodologies such as few-shot prompting holds the potential to yield highly flexible and engaging natural language generation models. In few-shot translation, the prompt includes a selection of translation examples (as depicted in Table 7.1), distinguishing it from the one-shot prompt solely by this inclusion.

7.3 Baseline System

In our investigation, we employed a baseline system constructed by amalgamating three distinct datasets for the Algerian Arabic dialect to Modern Standard Arabic (MSA) machine translation task: MADAR [152], PADIC [153], and our in-house dataset, referred to as ANMaT. Table 5.2 provides an overview of the statistics for each dataset, encompassing the number of parallel sentences, total word count, vocabulary size, and average sentence length. From the MADAR corpus, which encompasses 25 Arabic dialects, we selected 2,000 sentences representing the Algerian Arabic dialect. From PADIC, which encompasses dialects from five regions including Algeria (Algiers and Annaba), Tunisia, Syria, and Palestine, we extracted a pair of 6,412 sentences representing the Algerian dialect. Lastly, from our in-house dataset, meticulously curated to facilitate machine translation, we gathered 1,800 bilingual sentence pairs meticulously curated by two proficient native speakers of Algerian Arabic Dialect (DZDA) and Modern Standard Arabic (MSA). The consolidated corpus comprises a total of 10,212 sentence pairs across all three datasets.

7.4 Experiments and Evaluation

We strategically employed ChatGPT, a state-of-the-art (SOTA) language model, to explore its capabilities in translating between Modern Standard Arabic (MSA) and Algerian Darja (DZDA), an under-represented language pair in machine translation research (refer to Fig-

ure 7.2). Our investigation centered on assessing the model's adaptability and effectiveness under two distinct learning paradigms: zero-shot and few-shot scenarios.

7.4.1 The Scenario of zero-shot Prompting

In the zero-shot scenario, ChatGPT was engaged without being exposed to any specific training examples for the MSA-DZDA translation task. For this experimental condition, we relied solely on a standard prompt, the specifics of which are detailed in Table 7.1, to input translation requests to the model. This scenario was particularly revealing as it provided insights into the innate abilities of ChatGPT to handle translation tasks relying solely on its pre-trained knowledge base, thus assessing the generalizability of the model across linguistically and culturally distinct language pairs without prior task-specific adaptation.

7.4.2 The Scenario of few-shot Prompting

Conversely, the few-shot scenario was designed to examine the efficacy of incremental learning, where ChatGPT was primed with a modest dataset comprising just 50 sentence pairs in MSA and DZDA. Despite the limited data input, this approach was intended to observe how even minimal task-specific training could potentially enhance the model's translation performance. Importantly, we maintained the default parameters of ChatGPT throughout this scenario to eliminate any confounding variables that could impact the interpretability of the results, ensuring that any observed performance differences could be attributed directly to the influence of the few-shot learning method.

These experimental setups allowed us to conduct a nuanced examination of how variations in prompt design and the introduction of minimal targeted training influence the performance of a leading-edge neural language model in a machine translation task. This, in turn, provides valuable insights into the scalability and flexibility of few-shot learning within the realm of neural machine translation, particularly for language pairs like MSA and DZDA, which are typically less represented in computational linguistic resources. The findings of this study are expected to contribute significantly to the ongoing discourse on optimizing machine translation systems for a broader array of languages using advanced language models such as ChatGPT.

7.4.3 Evaluation metrics

For the DZDA-MSA language pair, we investigated the performance of both the NMT model based on the GRU-attentional-based Seq2Seq architecture and ChatGPT model using the following metrics: **BLEU**, **ChrF** and **COMET**.

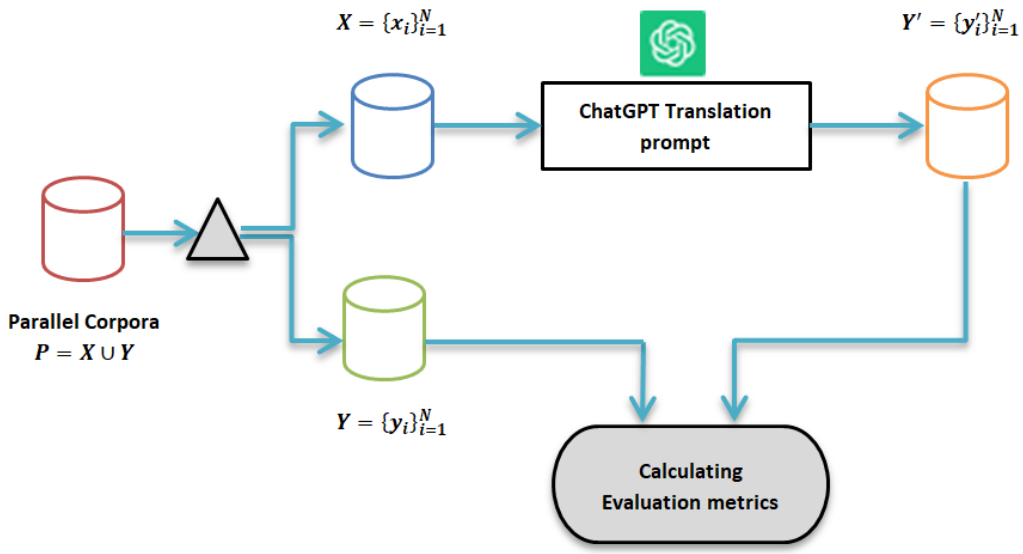

Figure 7.2: The ChatGPT Translation approach

The shared task on translation metrics, as suggested by [167], advocates for the adoption of neural network-based metrics due to their strong correlation with human evaluations and their resilience to domain shifts. While we include findings utilizing BLEU and ChrF as benchmarks, our primary focus lies in assessing performance through the model-centric metric COMET.

7.5 Results and Discussions

The assessment outcomes provide insights into the translation proficiency of the Seq2Seq and ChatGPT models in converting text between DZDA (Algerian Dialectal Arabic) and MSA (Modern Standard Arabic), as determined by key evaluation metrics.

7.5.1 MSA-to-DZA Translation

The translation evaluation outcomes for the language pair MSA→DZDA are displayed in Table 7.2. The Seq2Seq model attained scores of $COMET = 73.0$, $BLEU = 17.7$, and $ChrF = 41.5$. In contrast, the ChatGPT model achieved higher scores in the zero-shot scenario, with $COMET = 75.3$, $BLEU = 19.9$, $andChrF = 60.4$. Furthermore, in the few-shot scenario, the ChatGPT model demonstrated further enhancement, with scores of $COMET = 75.4$, $BLEU = 19.9$, and $ChrF = 60.7$.

These results indicate that the ChatGPT model surpassed the Seq2Seq model across all evaluation metrics. Notably, the translation quality for the MSA→DZDA language pair

exhibited considerable enhancement across all models. When comparing the zero-shot and few-shot scenarios of ChatGPT, we notice marginal improvements in COMET, BLEU, and ChrF scores. While these improvements are not substantial, they suggest that integrating a few training examples has a positive impact on the model's translation capabilities.

In summary, the ChatGPT model performs better than the Seq2Seq model in translating MSA→-DZDA. The language pair is well translated, and the ChatGPT model benefits from the few-shot scenario to slightly enhance translation quality.

Table 7.2: MSA→DZDA and DZDA→MSA translation performance

	MSA→DZDA			DZDA→MSA		
System	COMET	BLEU	ChrF	COMET	BLEU	ChrF
Seq2Seq	73.0	17.7	41.5	73.5	19.1	45.4
ChatGPT (0-shot)	75.3	**19.9**	60.4	76.4	21.2	53.5
ChatGPT (F-shot)	**75.4**	**19.9**	**60.7**	**76.6**	**21.3**	**53.8**

7.5.2 DZDA-to-MSA Translation

However, the assessment outcomes for the translation from DZDA to MSA, outlined in Table 7.2, reveal the following: the Seq2Seq model attained scores of $COMET = 73.5$, $BLEU = 19.1$, and $ChrF = 45.4$. In contrast, the ChatGPT model performed better in the zero-shot scenario, achieving scores of $COMET = 76.4$, $BLEU = 21.2$, and $ChrF = 53.5$. Moreover, the few-shot scenario with ChatGPT resulted in further enhancements, yielding scores of $COMET = 76.6$, $BLEU = 21.3$, and $ChrF = 53.8$.

These findings indicate that once again, the ChatGPT model outperforms the Seq2Seq model for the task of translating from DZDA to MSA, with higher COMET, BLEU, and ChrF scores. This underscores ChatGPT's efficacy in translating from the Algerian Arabic dialect (DZDA) to Modern Standard Arabic (MSA). Overall, the DZDA-to-MSA language pair exhibits satisfactory translation quality across all models.

When comparing these outcomes with those previously discussed for MSA-to-DZDA translation, we notice consistent performance from ChatGPT across both language directions. In both the zero-shot and few-shot scenarios, ChatGPT demonstrates enhanced translation quality compared to the Seq2Seq model. It is important to acknowledge that while the specific metric scores may vary slightly between the two language pairs, the overall trend remains consistent. These results affirm the effectiveness of ChatGPT in improving translation capabilities for low-resource language pairs like MSA and DZDA, regardless of the translation direction.

Transitioning from zero-shot to few-shot scenarios, ChatGPT models display improved translation performance, evidenced by higher COMET, BLEU, and ChrF scores across both language pairs. Introducing additional training instances through the few-shot scenario enhances the models' translation capacity, albeit to a modest extent.

Language models such as BERT and GPT, pre-trained on vast text corpora, offer considerable potential for enhancing translation quality in low-resource and minority languages. Fine-tuning these models on particular language pairs with restricted datasets can substantially improve translation quality, thereby fostering advancements in education, communication, and economic development within minority language communities encountering resource limitations.

7.5.3 Human analysis

In our evaluation, ChatGPT's translations from MSA to DZDA often resembled Moroccan Arabic Dialect rather than Algerian Arabic Dialect, suggesting a bias possibly due to exposure to Moroccan Arabic during pre-training. This bias could be due to the prevalence of Moroccan Arabic data in the training corpus or the impact of internet content that predominantly favours Moroccan sources.

Nevertheless, In the western region of Algeria, the dialect shares similarities with Moroccan, contributing to linguistic overlap. However, after multiple response regenerations, translations tended to adopt a more Algerian-like style (see Table B.1). This indicates that ChatGPT, despite initially favoring Moroccan Arabic, was able to adapt and incorporate Algerian Arabic characteristics after multiple iterations. In addition to the previously discussed issue, the Seq2Seq NMT model employed for DZDA→MSA translation produced inaccurate translations, as observed in the given examples (Table A.1). It incorrectly used the term المحطة (the station) instead of the intended words السكَايب (Skype). This problem arises from the model's limitations in capturing precise semantic nuances and distinguishing between multiple word meanings. To tackle this issue, various measures can be implemented such as:

- Enhancing the training data with varied examples and integrating context-specific annotations can enhance accuracy.

- Refining the model with domain-specific data or using advanced techniques, like integrating lexical resources or utilizing contextual embeddings, could improve translation quality.

- Regular model evaluation and iterative refinement based on feedback and error analysis can also contribute to rectifying such inaccuracies.

By leveraging these strategies, we can work towards improving the translation performance of the Seq2Seq NMT model for DZDA-MSA language pair.

The results underscore the challenges of training language models on dialectal variations within a language. More research is necessary to comprehend the factors influencing model biases and to develop strategies for accurately representing the diverse linguistic features of Algerian Arabic in future model iterations.

7.6 Summary

In conclusion, this chapter has provided an extensive exploration of the comparative landscape between Seq2Seq Neural Machine Translation (NMT) models and GPT-based models, shedding light on their architectural nuances and performance characteristics. We meticulously dissected the architectural intricacies of both approaches, highlighting the distinctive features of Seq2Seq models with encoder-decoder frameworks and attention mechanisms, as well as GPT-based models with pre-trained transformer architecture and autoregressive generation capabilities. Through the implementation of a robust experimental framework, including zero-shot and few-shot prompting scenarios, we evaluated the performance of both model types across various translation tasks. Leveraging comprehensive evaluation metrics such as COMET, BLEU, and ChrF scores, we conducted a rigorous analysis of translation quality and effectiveness. The findings revealed nuanced insights into the strengths and limitations of each approach, further augmented by a qualitative human analysis component.

Our meticulous analysis of the DZDA→MSA and MSA→DZDA translation tasks revealed valuable insights into the strengths and limitations of each approach. Notably, the ChatGPT model consistently outperformed the Seq2Seq model in both translation directions, demonstrating higher COMET, BLEU, and ChrF scores. Specifically, in the MSA→DZDA translation, the ChatGPT model showcased superior performance, benefiting from the few-shot scenario to slightly enhance translation quality. Conversely, in the DZDA→MSA translation, the ChatGPT model's effectiveness was evident, underscoring its efficacy in translating from the Algerian Arabic dialect (DZDA) to Modern Standard Arabic (MSA). Overall, the translation quality for both language pairs was satisfactory across all models. Additionally,

transitioning from zero-shot to few-shot scenarios yielded improved translation performance for ChatGPT models, as evidenced by higher COMET, BLEU, and ChrF scores across both language pairs. These findings contribute to a deeper understanding of the dynamics between Seq2Seq and GPT-based models in the domain of machine translation, laying the groundwork for future advancements and innovations in the field.

Chapter 8

Conclusion

8.1 Summary

The success of machine translation heavily relies on the availability of substantial and high-quality data, particularly parallel corpora, which are essential for training competitive Neural Machine Translation (NMT) models. In our study, we focused on the Algerian dialectal Arabic (DZDA) language pair, chosen due to its under-representation in mainstream NMT despite its relevance in media content. By collecting, preprocessing, and aligning DZDA-Modern Standard Arabic (MSA) parallel sentences from diverse sources, we tackled the significant challenge of data scarcity in low-resource languages. To address this issue, we proposed and implemented various data augmentation methods.

One such approach involved the development of a novel bilingual corpus, ANMaT, sourced from multiple social media platforms. This corpus, comprising approximately 1,800 DZDA-MSA sentence pairs, served as the foundation for our NMT system designed specifically for low-resource languages. Leveraging a Seq2Seq architecture, we further optimized the system's performance by adjusting hyper-parameters.

The evaluation of the augmented and consolidated corpora, in comparison with baseline systems employing different word segmentation methods, was performed using statistical significance tests and metrics such as BLEU, ChrF, and COMET. The findings from this evaluation study highlight two critical aspects. Firstly, data augmentation strategies combined with subword segmentation significantly improved the quality of NMT systems designed for dialectal Arabic languages in low-resource contexts, effectively addressing our first research question (**RQ1**). Moreover, we investigated the suitability of GPT-based models, particularly ChatGPT, as an alternative for machine translation in low-resource settings, addressing our second research question (**RQ2**). Our experiments revealed that GPT-based models, including ChatGPT, demonstrated promising results in low-data conditions compared to both

the Seq2Seq NMT model and commercial MT systems. These findings underscore the potential of leveraging advanced language models for addressing the challenges of translating low-resource languages effectively.

8.2 Limitations

In conclusion, while our computational resources constrained us from directly comparing combined augmented corpora on the Seq2Seq model, we believe that our approach serves as a significant contribution in its own right. By meticulously outlining our methodology, we provide a framework for understanding the nuances behind the observed variations in performance. This transparency not only enhances the reproducibility of our findings but also offers valuable insights into the intricate dynamics of data augmentation techniques in low-resource scenarios.

8.3 Future work

In our efforts, we advocate for expanding the DZDA-MSA parallel corpus size by incorporating texts sourced from other meticulously curated repositories. Augmenting the corpus not only amplifies its size but also elevates its quality by rectifying grammatical inaccuracies. Additionally, while our study did not entail disambiguating the part-of-speech (POS) of words within sentences, we recognize the significance of this aspect. Given that word segmentation varies depending on their POS and contextual usage, we strongly advocate for incorporating POS disambiguation in future research endeavors. Furthermore, leveraging human evaluation or expert judgment facilitates a more nuanced analysis of NMT models, providing valuable insights into their performance. Looking ahead, we envision conducting our research on a broader scale to delve deeper into the intricacies of low-resource MT scenarios. Moreover, exploring alternative Large Language Models (LLMs) for MT tasks in such scenarios presents an intriguing avenue for future investigation.

Bibliography

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[2] Usama Khalid, Mirza Omer Beg, and Muhammad Umair Arshad. Rubert: A bilingual roman urdu bert using cross lingual transfer learning. *arXiv preprint arXiv:2102.11278*, 2021.

[3] Mingye Wang, Pan Xie, Yao Du, and Xiaohui Hu. T5-based model for abstractive summarization: A semi-supervised learning approach with consistency loss functions. *Applied Sciences*, 13:7111, 06 2023.

[4] Miguel Domingo, Mercedes Garcıa-Martınez, Alexandre Helle, Francisco Casacuberta, and Manuel Herranz. How much does tokenization affect neural machine translation? *arXiv preprint arXiv:1812.08621*, 2019.

[5] Andr´s Kornai. Digital language death. *PLOS ONE*, 8(10):1–11, 10 2013.

[6] Bonnie J Dorr, Pamela W Jordan, and John W Benoit. A survey of current paradigms in machine translation. In *Advances in computers*, volume 49, pages 1–68. Elsevier, 1999.

[7] Martin Haspelmath. Pre-established categories don't exist: Consequences for language description and typology. *Linguistic Typology*, 11(1):119–132, 2007.

[8] Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. Survey of Low-Resource Machine Translation. *Computational Linguistics*, 48(3):673–732, 09 2022.

[9] Philipp Koehn and Rebecca Knowles. Six challenges for neural machine translation. In Thang Luong, Alexandra Birch, Graham Neubig, and Andrew Finch, editors, *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver, aug 2017. Association for Computational Linguistics.

[10] Sainik Kumar Mahata, Soumil Mandal, Dipankar Das, and Sivaji Bandyopadhyay. SMT vs NMT: A comparison over Hindi and Bengali simple sentences. In Gurpreet Singh Lehal, Dipti Misra Sharma, and Rajeev Sangal, editors, *Proceedings of the 15th International Conference on Natural Language Processing*, pages 175–182, International Institute of Information Technology, Hyderabad, India, dec 2018. NLP Association of India.

[11] Maria Stasimioti, Vilelmini Sosoni, Katia Kermanidis, and Despoina Mouratidis. Machine translation quality: A comparative evaluation of SMT, NMT and tailored-NMT outputs. In André Martins, Helena Moniz, Sara Fumega, Bruno Martins, Fernando Batista, Luisa Coheur, Carla Parra, Isabel Trancoso, Marco Turchi, Arianna Bisazza, Joss Moorkens, Ana Guerberof, Mary Nurminen, Lena Marg, and Mikel L. Forcada, editors, *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 441–450, Lisboa, Portugal, November 2020. European Association for Machine Translation.

[12] Claudia Matos Veliz, Orphée De Clercq, and Veronique Hoste. Is neural always better? smt versus nmt for dutch text normalization. *Expert Systems with Applications*, 170:114500, 2021.

[13] Ken Peffers, Tuure Tuunanen, Marcus Rothenberger, and S. Chatterjee. A design science research methodology for information systems research. *Journal of Management Information Systems*, 24:45–77, 01 2007.

[14] Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Phrase-based & neural unsupervised machine translation. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049, Brussels, Belgium, oct – nov 2018. Association for Computational Linguistics.

[15] Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor O.K. Li. Universal neural machine translation for extremely low resource languages. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 344–354, New Orleans, Louisiana, jun 2018. Association for Computational Linguistics.

[16] Rico Sennrich and Biao Zhang. Revisiting low-resource neural machine translation: A case study. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings*

of the 57th Annual Meeting of the Association for Computational Linguistics, pages 211–221, Florence, Italy, jul 2019. Association for Computational Linguistics.

[17] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, Aug 2016. Association for Computational Linguistics.

[18] Baligh Babaali and Mohammed Salem. Survey of the Arabic Machine Translation Corpora. *Lecture Notes in Networks and Systems*, 593 LNNS:205–219, 2023.

[19] Baligh Babaali and Mohammed Salem. Arabic machine translation: A panoramic survey. *Available at SSRN 4312742*, 2022.

[20] Baligh Babaali, Mohammed Salem, and Nawaf R. Alharbe. Breaking language barriers with ChatGPT: enhancing low-resource machine translation between algerian arabic and MSA. *International Journal of Information Technology*, May 2024.

[21] Mirko Plitt and François Masselot. A productivity test of statistical machine translation post-editing in a typical localisation context. In *Prague Bulletin of Mathematical Linguistics*, 2010.

[22] Ebtesam H. Almansor and Ahmed Al-Ani. A hybrid neural machine translation technique for translating low resource languages. In Petra Perner, editor, *Machine Learning and Data Mining in Pattern Recognition*, pages 347–356, Cham, 2018. Springer International Publishing.

[23] Arwa Alqudsi, Nazlia Omar, and Khalid Shaker. Arabic machine translation: a survey. *Artificial Intelligence Review*, 42(4):549–572, 2014.

[24] A Fassi Fehri. *Issues in the structure of Arabic clauses and words*, volume 29. Springer Science & Business Media, 1993.

[25] Achraf Chalabi. Mt-based transparent arabization of the internet tarjim. com. In *Conference of the Association for Machine Translation in the Americas*, pages 189–191. Springer, 2000.

[26] Kevin Daimi. Identifying syntactic ambiguities in single-parse arabic sentence. *Computers and the Humanities*, 35(3):333–349, 2001.

[27] Valia Kordoni and Iliana Simova. Multiword expressions in machine translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1208–1211, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA).

[28] Andrea Zaninello and Alexandra Birch. Multiword expression aware neural machine translation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3816–3825, Marseille, France, May 2020. European Language Resources Association.

[29] Danah m. Boyd and Nicole B. Ellison. Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1):210–230, 10 2007.

[30] Muhammad Zafar Yaqub and Abdullah Alsabban. Knowledge sharing through social media platforms in the silicon age. *Sustainability*, 15(8), 2023.

[31] Anamaria Dutceac Segesten, Michael Bossetta, Nils Holmberg, and Diederick Niehorster. The cueing power of comments on social media: how disagreement in facebook comments affects user engagement with news. *Information, Communication & Society*, 25(8):1115–1134, 2022.

[32] Sung-joon Yoon. Does social capital affect sns usage? a look at the roles of subjective well-being and social identity. *Computers in Human Behavior*, 41:295–303, 2014.

[33] Richard Rogers. Visual media analysis for instagram and other online platforms. *Big Data & Society*, 8(1):20539517211022370, 2021.

[34] Yingdan Lu and Yilang Peng. The mobilizing power of visual media across stages of social-mediated protests. *Political Communication*, 0(0):1–28, 2024.

[35] A. Lenhart, R. Ling, S. Campbell, and A. Elder. Teens, social media, & technology overview 2015. https://www.pewresearch.org/internet/2015/04/09/teens-social-media-technology-2015/.

[36] Natalie Pang and Yue Ting Woo. What about whatsapp? a systematic review of whatsapp and its role in civic and political engagement. *First Monday*, 25(12), Jan. 2020.

[37] Nicole B. Ellison, Charles Steinfield, and Cliff Lampe. The benefits of facebook "friends:" social capital and college students' use of online social network sites. *Journal of Computer-Mediated Communication*, 12(4):1143–1168, 2007.

[38] Joanna Davis, Hans-Georg Wolff, Monica L. Forret, and Sherry E. Sullivan. Networking via linkedin: An examination of usage and career benefits. *Journal of Vocational Behavior*, 118:103396, 2020.

[39] Diego Moussallem, Matthias Wauer, and Axel Cyrille Ngonga Ngomo. Semantic web for machine translation: Challenges and directions. *CEUR Workshop Proceedings*, 2576:1–9, 2019.

[40] Kareem Darwish. Arabizi detection and conversion to arabic. In *Proceedings ofthe EMNLP 2014 Workshop on Arabic Natural Langauge Processing (ANLP)*, pages 217–224, Doha, Qatar, 2015.

[41] M. D. Okpor. Machine translation approaches: Issues and challenges. *IJCSI International Journal of Computer Science Issues*, 11(2):159–165, Sep 2014.

[42] Neeha Ashraf and Manzoor Ahmad. Machine translation techniques and their comparative study. *International Journal of Computer Applications*, 125(7):25–31, Sep 2015. Published by Foundation of Computer Science (FCS), NY, USA.

[43] Thi-Ngoc-Diep DO. *Extraction de corpus parallèle pour la traduction automatique depuis et vers une langue peu dotée*. PhD thesis, UNIVERSITÉ DE GRENOBLE, 2011.

[44] Makoto Nagao. Framework of a mechanical translation between japanese and english by analogy principle. *Artificial and Human Intelligence, A. Elithorn and R. Banerji (eds.) North-Holland*, pages 173–180, 1984.

[45] Peter F Brown, John Cocke, Stephen A Della Pietra, Vincent J Della Pietra, Frederick Jelinek, Robert L Mercer, and Paul Roossin. A statistical approach to language translation. In *Coling Budapest 1988 Volume 1: International Conference on Computational Linguistics*, 1988.

[46] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, 1993.

[47] Philipp Koehn. *Neural machine translation*. Cambridge University Press, 2020.

[48] Mirjam Sepesy Maučec and Gregor Donaj. Machine translation and the evaluation of its quality. In *Recent Trends in Computational Intelligence*. IntechOpen, 2019.

[49] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, Jul 2002. Association for Computational Linguistics.

[50] Maja Popović. chrF: character n-gram F-score for automatic MT evaluation. In Ondřej Bojar, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara Logacheva, and Pavel Pecina, editors, *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, sep 2015. ACL.

[51] Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. COMET: A Neural Framework for MT Evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Stroudsburg, PA, USA, 2020. ACL.

[52] Ahmad T Al-Taani and Zeyad M Hailat. A direct english-arabic machine translation system. *Information Technology Journal*, 4(3):256–261, 2005.

[53] Abraham Ittycheriah and Salim Roukos. Direct translation model 2. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 57–64, Rochester, New York, Apr 2007. Association for Computational Linguistics.

[54] Chafia Mankai and Ali Mili. Machine translation from arabic to english and french. *Information Sciences-Applications*, 3(2):91–109, 1995.

[55] Yasser Salem, Arnold Hensman, and Brian Nolan. Implementing arabic-to-english machine translation using the role and reference grammar linguistic model. In *Proceedings of the Eighth Annual International Conference on Information Technology and Telecommunication*, pages 103–110, 2008.

[56] ThuyLinh Nguyen and Stephan Vogel. Context-based Arabic morphological analysis for machine translation. In *CoNLL 2008: Proceedings of the Twelfth Conference on Computational Natural Language Learning*, pages 135–142, Manchester, England, Aug 2008. Coling 2008 Organizing Committee.

[57] Doaa Samy and Ana González-Ledesma. Pragmatic annotation of discourse markers in a multilingual parallel corpus (arabic-spanish-english). In *LREC*, 2008.

[58] M. M. Abu Shquier and T. M. T. Sembok. Word agreement and ordering in english-arabic machine translation. In *2008 International Symposium on Information Technology*, volume 1, pages 1–10, 2008.

[59] Jakob Elming and Nizar Habash. Syntactic reordering for English-Arabic phrase-based machine translation. In *Proceedings of the EACL 2009 Workshop on Computational Approaches to Semitic Languages*, pages 69–77, Athens, Greece, Mar 2009. Association for Computational Linguistics.

[60] Romaric Besançon, Djamel Mostefa, Ismaïl Timimi, Stéphane Chaudiron, Mariama Laïb, and Khalid Choukri. Arabic, english and french: three languages in a filtering systems evaluation project. *MEDAR*, pages 163–167, 2009.

[61] Wael Salloum and Nizar Habash. Elissa: A dialectal to standard arabic machine translation system. In *COLING 2012: Demonstration Papers*, December 2012, pages 385–392, Mumbai, 2012.

[62] Mohamed Ali Sghaier and Mounir Zrigui. Rule-based machine translation from tunisian dialect to modern standard arabic. *Procedia Computer Science*, 176:310–319, 2020.

[63] Abdelhadi Soudi, Violetta Cavalli-Sforza, and Abderrahim Jamari. Prototype english-to-arabic interlingua-based mt system. In *Proceedings of the Third International Conference on Language Resources and Evaluation: Workshop on Arabic Language Resources and Evaluation: Status and Prospects*, pages 18–25, Las Palmas de Gran Canaria, Spain, 01 2002.

[64] Khaled Shaalan, Azza Abdel Monem, Ahmed Rafea, and Hoda Baraka. Mapping interlingua representations to feature structures of arabic sentences. In *The Challenge of Arabic for NLP/MT. International Conference at the British Computer Society, London*, pages 149–159, 2006.

[65] Pierrette Bouillon, Ismahene Sonia Halimi Mallem, Yukie Nakao, Kyoko Kanzaki, Hitoshi Isahara, Nikolaos Tsourakis, Marianne Starlander, Beth Ann Hockey, and Emmanuel Rayner. Developing non-european translation pairs in a medium-vocabulary medical speech translation system. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC)*, pages 1741–1748, 2008.

[66] Sameh Alansary. Interlingua-based machine translation systems: Unl versus other interlinguas. *The Egyptian Journal of Language Engineering*, 1, Jan 2014.

[67] M Attia. Implications of the agreement features in machine translation. *Al-Azhar University*, 2002.

[68] Khaled Shaalan, Ahmed Rafea, Azza Abdel Moneim, and Hoda Baraka. Machine translation of english noun phrases into arabic. *International Journal of Computer Processing of Oriental Languages*, 17(02):121–134, 2004.

[69] Omar Shirko, Nazlia Omar, Haslina Arshad, and Mohammed Albared. Machine translation of noun phrases from arabic to english using transfer-based approach. *Journal of Computer Science*, 6(3):350, 2010.

[70] Khaled Shaalan, Ashraf Hendam, and Ahmed Rafea. An English-Arabic bi-directional machine translation tool in the agriculture domain: A rule-based transfer approach for translating expert systems. *IFIP Advances in Information and Communication Technology*, pages 281–290, 2010.

[71] Arwa Hatem, Nazlia Omar, and Khalid Shaker. Morphological analysis for rule based machine translation. In *2011 International Conference on Semantic Technology and Information Retrieval*, pages 260–263. IEEE, 2011.

[72] Mathieu Guidere. Toward corpus-based machine translation for standard arabic. *Translation Journal*, 6(1), 2002.

[73] Kfir Bar and N. Dershowitz. Using verb paraphrases for arabic-to-english example-based translation. *Machine Translation and Morphologically-rich Languages*, 2011.

[74] Kfir Bar and N. Dershowitz. *Using semantic equivalents for Arabic-to-English: Example-based translation*, pages 49–72. John Benjamins Publishing, Amsterdam, Netherlands, 2012.

[75] Violetta Cavalli-Sforza and Aaron Phillips. *Using morphology to improve Example-Based Machine Translation*, pages 23–48. John Benjamins Publishing, Jan 2012.

[76] T El-Shishtawy and A El-Sammak. The best templates match technique for example based machine translation. *arXiv preprint arXiv:1406.1241*, 2014.

[77] Sasa Hasan, Anas El Isbihani, and Hermann Ney. Creating a large-scale arabic to french statistical machine translation system. In *LREC*, pages 855–858, 2006.

[78] Mona Diab, Mahmoud Ghoneim, and Nizar Habash. Arabic diacritization in the context of statistical machine translation. In *Proceedings of MT-Summit*, 2007.

[79] Ruhi Sarikaya, Yonggang Deng, and Yuqing Gao. Context dependent word modeling for statistical machine translation using part-of-speech tags. In *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[80] Ibrahim Badr, Rabih Zbib, and James Glass. Syntactic phrase reordering for English-to-Arabic statistical machine translation. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 86–93, Athens, Greece, Mar 2009. Association for Computational Linguistics.

[81] Nizar Habash and Jun Hu. Improving Arabic-Chinese statistical machine translation using English as pivot language. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 173–181, Athens, Greece, Mar 2009. Association for Computational Linguistics.

[82] Marine Carpuat, Yuval Marton, and Nizar Habash. Improving Arabic-to-English statistical machine translation by reordering post-verbal subjects for alignment. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 178–183, Uppsala, Sweden, Jul 2010. Association for Computational Linguistics.

[83] Mossa Ghurab, Yueting Zhuang, Jiangqin Wu, and Maan Younis Abdullah. Arabic-chinese and chinese-arabic phrase-based statistical machine translation systems. *Inf. Technol. J*, 9(4):666–672, 2010.

[84] Arianna Bisazza and Marcello Federico. Chunk-based verb reordering in VSO sentences for Arabic-English statistical machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 235–243, Uppsala, Sweden, Jul 2010. Association for Computational Linguistics.

[85] Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and Stephan Vogel. Qcri machine translation systems for iwslt 16. *arXiv preprint arXiv:1701.03924*, 2017.

[86] Fatma Mallek, Billal Belainine, and Fatiha Sadat. Arabic social media analysis and translation. *Procedia Computer Science*, 117:298–303, 2017.

[87] Sara Ebrahim, Doaa Hegazy, Mostafa Gadal Haqq M. Mostafa, and Samhaa R. El-Beltagy. Detecting and integrating multiword expression into english-arabic statistical machine translation. *Procedia Computer Science*, 117:111–118, 2017.

[88] Fares Aqlan, Xiaoping Fan, Abdullah Alqwbani, and Akram Al-Mansoub. Improved arabic-chinese machine translation with linguistic input features. *Future Internet*, 11:22, 2019.

[89] Haytham Alsharaf, Sylviane Cardey, and Peter Greenfield. French to arabic machine translation: the specificity of language couples. In *Proc. of the 9th Annual Workshop of the European Association for Machine Translation (EAMT), Malta, April*, 2004.

[90] Kristina Toutanova, Hisami Suzuki, and Achim Ruopp. Applying morphology generation models to machine translation. In *Proceedings of ACL-08: HLT*, pages 514–522, Columbus, Ohio, Jun 2008. Association for Computational Linguistics.

[91] Ahmed Hatem and Amin Nassar. Modified dijstra-like search algorithm for english to arabic machine translation system. *Proceedings EAMT*, 2008:12th, 2008.

[92] Evgeny Matusov, Gregor Leusch, and Hermann Ney. Learning to Combine Machine Translation Systems. In *Learning Machine Translation*. The MIT Press, nov 2008. _eprint: https://academic.oup.com/mit-press-scholarship-online/book/0/chapter/245133702/chapter-ag-pdf/44575073/book_29423_section_245133702.ag.pdf.

[93] Nizar Habash, B. Dorr, and Christof Monz. Symbolic-to-statistical hybridization: extending generation-heavy machine translation. *Machine Translation*, 23:23–63, 2009.

[94] Rasha Al Dam and Ahmed Guessoum. Building a neural network-based english-to-arabic transfer module from an unrestricted domain. In *2010 International Conference on Machine and Web Intelligence*, pages 94–101, 2010.

[95] Hassan Sawaf. Arabic dialect handling in hybrid machine translation. In *AMTA 2010 - 9th Conference of the Association for Machine Translation in the Americas*, 2010.

[96] Mouiad Alawneh and Tengku Mohd Sembok. Handling agreement and words reordering in machine translation from english to arabic using hybrid-based systems. *Journal of Computer Science*, 11(6):93–97, 2011.

[97] Khaled Shaalan and Ahmad Hany Hossny. Automatic rule induction in arabic to english machine translation framework. *Challenges for Arabic Machine Translation*, 9(2012):135, 2012.

[98] Marwan Akeel and Ravi Mishra. Ann and rule based method for english to arabic machine translation. *Int. Arab J. Inf. Technol.*, 11(4):396–405, 2014.

[99] Emad Mohamed and Fatiha Sadat. Hybrid arabic–french machine translation using syntactic re-ordering and morphological pre-processing. *Computer Speech & Language*, 32(1):135–144, 2015.

[100] Rached Zantout and Ahmed Guessoum. Obstacles facing arabic machine translation: building a neural network-based transfer module. *Papers in Translation Studies*, pages 229–253, 2015.

[101] Jeffrey L. Elman. Finding structure in time. *Cognitive Science*, 14(2):179–211, 1990.

[102] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR'15)*, 2015.

[103] M. Schuster and K.K. Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.

[104] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[105] Amjad Almahairi, Kyunghyun Cho, Nizar Habash, and Aaron Courville. First result on arabic neural machine translation. *arXiv preprint arXiv:1606.02680*, 2016.

[106] Francisco Guzmán, Houda Bouamor, Ramy Baly, and Nizar Habash. Machine translation evaluation for Arabic using morphologically-enriched embeddings. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1398–1408, Osaka, Japan, dec 2016. The COLING 2016 Organizing Committee.

[107] Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872, Vancouver, Canada, Jul 2017. Association for Computational Linguistics.

[108] Gyu-Hyeon Choi, Jong-Hun Shin, and Young-Kil Kim. Improving a multi-source neural machine translation model with corpus extension for low-resource languages. *arXiv Preprint:1709.08898*, 2017.

[109] Pamela Shapiro and Kevin Duh. Morphological word embeddings for Arabic neural machine translation in low-resource setting. In *Proceedings of the Second Workshop on Subword/Character LEvel Models*, pages 1–11, New Orleans, Jun 2018. Association for Computational Linguistics.

[110] Abdullah Alrajeh. A recipe for arabic-english neural machine translation. *arXiv preprint arXiv:1808.06116*, 2018.

[111] Manar Alkhatib and Khaled Shaalan. Paraphrasing arabic metaphor with neural machine translation. *Procedia Computer Science*, 142:308–314, 2018.

[112] Laith H Baniata, Seyoung Park, and Seong-bae Park. A neural machine translation model for arabic dialects that utilizes multitask learning (mtl). *Computational Intelligence and Neuroscience*, 2018:1–10, 2018.

[113] Fares Aqlan, Xiaoping Fan, Abdullah Alqwbani, and Akram Al-Mansoub. Arabic–chinese neural machine translation: Romanized arabic as subword unit for arabic-sourced translation. *IEEE Access*, 7:133122–133135, 2019.

[114] Mai Oudah, Amjad Almahairi, and Nizar Habash. The impact of preprocessing on Arabic-English statistical and neural machine translation. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 214–221, Dublin, Ireland, Aug 2019. European Association for Machine Translation.

[115] Mohamed Seghir Hadj Ameur, Ahmed Guessoum, and Farid Meziane. Improving arabic neural machine translation via n-best list re-ranking. *Machine Translation*, 33(4):279–314, 2019.

[116] Ibrahim Gashaw and HL Shashirekha. Amharic-arabic neural machine translation. *arXiv preprint arXiv:1912.13161*, 2019.

[117] Baijun Ji, Zhirui Zhang, Xiangyu Duan, Min Zhang, Boxing Chen, and Weihua Luo. Cross-lingual pre-training based transfer for zero-shot neural machine translation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:115–122, 04 2020.

[118] Nouhaila Bensalah, Habib Ayad, Abdellah Adib, and Abdelhamid Ibn El Farouk. Lstm vs. gru for arabic machine translation. In Ajith Abraham, Yukio Ohsawa, Niketa Gandhi, M.A. Jabbar, Abdelkrim Haqiq, Seán McLoone, and Biju Issac, editors, *Proceedings of the 12th International Conference on Soft Computing and Pattern Recognition (SoCPaR 2020)*, pages 156–165, Cham, 2021. Springer International Publishing.

[119] Safae Berrichi and Azzeddine Mazroui. Addressing limited vocabulary and long sentences constraints in english–arabic neural machine translation. *Arabian Journal for Science and Engineering*, 46(9):8245–8259, 2021.

[120] Youness Moukafih, Nada Sbihi, Mounir Ghogho, and Kamel Smaïli. Improving machine translation of arabic dialects through multi-task learning. In *20th International Conference Italian Association for Artificial Intelligence:AIxIA 2021*, pages 235–243, MILAN/Virtual, Italy, Dec 2021.

[121] El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. Investigating code-mixed modern standard arabic-egyptian to english machine translation. *arXiv preprint arXiv:2105.13573*, 2021.

[122] Emmanouil Stergiadis, Satendra Kumar, Fedor Kovalev, and Pavel Levin. Multi-domain adaptation in neural machine translation through multidimensional tagging. *arXiv preprint arXiv:a2102.10160v1*, 2021.

[123] Nouhaila Bensalah, Habib Ayad, Abdellah Adib, and Abdelhamid Ibn El Farouk. Cran: An hybrid cnn-rnn attention-based model for arabic machine translation. In Mohamed Ben Ahmed, Horia-Nicolai L. Teodorescu, Tomader Mazri, Parthasarathy Subashini, and Anouar Abdelhakim Boudhir, editors, *Networking, Intelligent Systems and Security*, pages 87–102, Singapore, 2022. Springer Singapore.

[124] Amel Slim, Ahlem Melouah, Usef Faghihi, and Khouloud Sahib. Improving neural machine translation for low resource algerian dialect by transductive transfer learning strategy. *Arabian Journal for Science and Engineering*, pages 1–8, 2022.

[125] Marwa Gaser, Manuel Mager, Injy Hamed, Nizar Habash, Slim Abdennadher, and Ngoc Thang Vu. Exploring segmentation approaches for neural machine translation of code-switched egyptian arabic-english text. *arXiv Preprint:2210.06990*, 2022.

[126] El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. Turjuman: A public toolkit for neural arabic machine translation. *arXiv preprint arXiv:2206.03933v1*, 2022.

[127] Diadeen Ali Hameed, Tahseen Ameen Faisal, Ali Mustafa Alshaykha, Ghanim Thiab Hasan, and Harith Abdullah Ali. Automatic evaluating of russian-arabic machine translation quality using meteor method. *AIP Conference Proceedings*, 2386(1):040036, 2022.

[128] Laith H. Baniata, Sangwoo Kang, and Isaac. K. E. Ampomah. A reverse positional encoding multi-head attention-based neural machine translation model for arabic dialects. *Mathematics*, 10(19), 2022.

[129] Haoran Devlin, Benjamin Van Durme, and Kenton Murray. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[130] Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual BERT? In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy, jul 2019. Association for Computational Linguistics.

[131] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, and other. Improving language understanding by generative pre-training, 2018.

[132] Steve Yang, Zulfikhar Ali, and Bryan Wong. Fluid-gpt (fast learning to understand and investigate dynamics with a generative pre-trained transformer): Efficient predictions of particle trajectories and erosion, 08 2023.

[133] Liyuan Liu, Jialu Liu, and Jiawei Han. Multi-head or single-head? an empirical comparison for transformer training. *arXiv preprint arXiv:2106.09650*, 2021.

[134] Thorsten Brants, Ashok Popat, Peng Xu, Franz Josef Och, and Jeffrey Dean. Large language models in machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 858–867, 2007.

[135] Yiheng Wang. Large language models evaluate machine translation via polishing. In *Proceedings of the 2023 6th International Conference on Algorithms, Computing and Artificial Intelligence*, ACAI '23, page 158–163, New York, NY, USA, 2024. Association for Computing Machinery.

[136] Chaoqun Liu, Wenxuan Zhang, Yiran Zhao, Anh Tuan Luu, and Lidong Bing. Is translation all you need? a study on solving multilingual tasks with large language models. *arXiv preprint arXiv:2403.10258*, 2024.

[137] Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Trans. Knowl. Discov. Data*, feb 2024.

[138] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

[139] Chenyang Lyu, Zefeng Du, Jitao Xu, Yitao Duan, Minghao Wu, Teresa Lynn, Alham Fikri Aji, Derek F. Wong, and Longyue Wang. A paradigm shift: The future of machine translation lies with large language models. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1339–1352, Torino, Italia, may 2024. ELRA and ICCL.

[140] Nooshin Pourkamali and Shler Ebrahim Sharifi. Machine translation with large language models: Prompt engineering for persian, english, and russian directions. *ArXiv*, abs/2401.08429, 2024.

[141] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *NIPS*, 2020-December, may 2020.

[142] Qingyu Lu, Baopu Qiu, Liang Ding, Liping Xie, and Dacheng Tao. Error analysis prompting enables human-like translation evaluation in large language models: A case study on chatgpt. *ArXiv*, abs/2303.13809, 2023.

[143] Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. Towards making the most of chatgpt for machine translation. *arXiv preprint arXiv:2303.13780*, 2023.

[144] Amr Hendy, Mohamed Gomaa Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. How good are gpt models at machine translation? a comprehensive evaluation. *ArXiv*, abs/2302.09210, 2023.

[145] Wenxiang Jiao, Wenxuan Wang, Jen tse Huang, Xing Wang, and Zhaopeng Tu. Is chatgpt a good translator? a preliminary study. *ArXiv*, abs/2301.08745, 2023.

[146] Yiming Chen, Yan Zhang, Chen Zhang, Grandee Lee, Ran Cheng, and Haizhou Li. Revisiting self-training for few-shot learning of language model. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9125–9135, Online and Punta Cana, Dominican Republic, nov 2021. ACL.

[147] Yao Dou, Maxwell Forbes, Rik Koncel-Kedziorski, Noah A. Smith, and Yejin Choi. Is GPT-3 text indistinguishable from human text? scarecrow: A framework for scrutinizing machine text. In *Proceedings of the 60th Annual Meeting of the ACL (Volume 1: Long Papers)*, pages 7250–7274, Dublin, Ireland, may 2022. ACL.

[148] Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. GPT3-to-plan: Extracting plans from text using GPT-3. *arXiv:2106.07131*, jun 2021.

[149] Zhaokun Jiang and Ziyin Zhang. Can chatgpt rival neural machine translation? a comparative study. *ArXiv*, abs/2401.05176, 2024.

[150] Karima Abidi and Kamel Smaïli. Creating multi-scripts sentiment analysis lexicons for algerian, moroccan and tunisian dialects. In *7th International Conference on Data Mining (DTMN 2021) Computer Science Conference Proceedings in Computer Science & Information Technology (CS & IT)*, Copenhagen, Denmark, Sep 2021.

[151] Maha J. Althobaiti. Automatic arabic dialect identification systems for written texts: A survey. *arXiv preprint arXiv:2009.12622*, 2020.

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

[153] Karima Meftouh, Salima Harrat, Salma Jamoussi, Mourad Abbas, and Kamel Smaili. Machine translation experiments on PADIC: A Parallel Arabic DIalect Corpus. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, pages 26–34, Shanghai, China, oct 2015.

[154] Younes Samih. *Dialectal Arabic processing using deep learning*. PhD thesis, Dissertation, Düsseldorf, Heinrich-Heine-Universität, 2017, 2017.

[155] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS'14)*, pages 3104–3112, 2014.

[156] Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. Understanding back-translation at scale. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural*

Language Processing, pages 489–500, Brussels, Belgium, oct – nov 2018. Association for Computational Linguistics.

[157] JiaJun Zhang and ChengQing Zong. Neural machine translation: Challenges, progress and future. *Science China Technological Sciences*, 63(10):2028–2050, October 2020.

[158] Thanh-Le Ha, Jan Niehues, and Alex Waibel. Toward multilingual neural machine translation with universal encoder and decoder. In Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, Rolando Cattoni, and Marcello Federico, editors, *Proceedings of the 13th International Conference on Spoken Language Translation*, Seattle, Washington D.C, dec 8-9 2016. International Workshop on Spoken Language Translation.

[159] Víctor M. Sánchez-Cartagena, Miquel Esplà-Gomis, Juan Antonio Pérez-Ortiz, and Felipe Sánchez-Martínez. Rethinking data augmentation for low-resource neural machine translation: A multi-task learning approach. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8502–8516, Online and Punta Cana, Dominican Republic, nov 2021. Association for Computational Linguistics.

[160] Nikolay Bogoychev and Rico Sennrich. Domain, translationese and noise in synthetic data for neural machine translation. *ArXiv*, abs/1911.03362, 2019.

[161] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. In Association for Computational Linguistics, editor, *Proceedings of the 8th Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST'14)*, pages 103–111, 2014.

[162] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In Sophia Ananiadou, editor, *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, jun 2007. Association for Computational Linguistics.

[163] Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. Opennmt: Open-source toolkit for neural machine translation. *arXiv preprint arXiv:1701.02810*, Mar 2017.

[164] Benoist Wolleb, Romain Silvestri, Giorgos Vernikos, Ljiljana Dolamic, and Andrei Popescu-Belis. Assessing the importance of frequency versus compositionality for subword-based tokenization in nmt. *arXiv preprint arXiv:2306.01393*, 2024.

[165] Matt Post. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the 3rd Conference on Machine Translation: Research Papers*, volume 1, pages 186–191, Stroudsburg, PA, USA, apr 2018. ACL.

[166] Mohd Javaid, Abid Haleem, Ravi Pratap Singh, Shahbaz Khan, and Ibrahim Haleem Khan. Unlocking the opportunities through chatgpt tool towards ameliorating the education system. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, 3(2):100115, 2023.

[167] Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F T Martins. Results of WMT22 Metrics Shared Task: Stop Using BLEU – Neural Metrics Are Better and More Robust. In *Proceedings of the 7th Conference on Machine Translation*, pages 46–68, Abu Dhabi, UAE (Hybrid), 2022. ACL.

Appendix A

Examples of Seq2Seq and ChatGPT translation

Table A.1: Examples of Seq2Seq and ChatGPT translation

DZDA→MSA	
Source (DZDA)	مَانيش سور بَالّي نقدروَا نكتبوهم في السكَايب
Reference (MSA)	لست متأكدَا من أنَّنَا نستطيع كتَابتهم في السكَايب
Seq2Seq (MSA)	لست متَاكدة من انَّنَا نستطيع كتَابتهم في المحطة
ChatGPT (MSA)	لست متأكدَاً ممَا إِذَا كنَا نستطيع كتَابتهم في سكَايب
Meaning	I am not sure that we can write them in Skype
Source (DZDA)	مَا نعرف ولَا يكفيني الوقت نزيد نمَاركي طَاليَانية
Reference (MSA)	لَا أعلم إِن كَان لدي الوقت لأَسجل لِلغة الَإيطَالية
Seq2Seq (MSA)	لَا أعلم إِن كَان لدي الوقت لأَسجل المحطة
ChatGPT (MSA)	لَا أعرف ولَا يكفيني الوقت لآتعلم الَإيطَالية
Meaning	I don't know if I have time to register for the Italian language
MSA→DZDA	
Source (MSA)	لم يعد جيم إِلَى المنزل بعد
Reference (DZDA)	جيم مَازَال مَا ولّا لدَّار
Seq2Seq (DZDA)	مَازَال جيم مَا لدَّار
ChatGPT (DZDA)	مَا رجعش جيم لّدَار حتَى الآن
Meaning	Jim hasn't come home yet
Source (MSA)	شرعت في قرَاءته لم أكمله بعد
Reference (DZDA)	أنَا بديت نقرَاه مَا زَال مَا كملتوش
Seq2Seq (DZDA)	أنَا بديت نقرَاه مَا كملتوش أومبعد
ChatGPT (DZDA)	بدأَت نقرَاه مَا كملتوش لحد الآن
Meaning	I started reading it, I haven't finished it yet

Appendix B

Examples of Dialectal Bias

Table B.1: Examples of Dialectal Bias

MSA source		لمَاذَا أنت آسف علَى شيء لم تفعله
DZDA target	1st response	علَاش كتسمحي رَاه تكوني متَأسفة علَى شيء مَا عملتيشه
	2nd response	علَاش كتَاسف علَى حَاجة مَا عملتهَاش
	3rd response	علَاش تطلب السمَاح لحَاجة مَا عملتهَاش
Meaning		Why are you sorry for something you didn't do
MSA source		أفهموني فيمَا نتحدث
DZDA target	1st response	اشكتلي علَى وَاش كنتي نقول
	2nd response	عفَاكم تفهموني فيمَا نتكلمو
	3rd response	فهموني فش تحكينَا فيه
	4th response	اش تفهمني وَاشنو كنتقول
	5th response	فهموني فَالي نحن نحكيوَا
Meaning		Let me understand what we are talking about
MSA source		إِنه يعرف مَا يريد
DZDA target	1st response	هو يعرف شنو يحب
	2nd response	هو يعرف شن يحب
	3rd response	يعرف وَاش يحب
Meaning		He knows what he wants