# Constructing Low Resource Approaches to Improve Speech-to-text Translation from Modern Standard Arabic to English

by

Rami Manna

S.B., Computer Science and Engineering
Massachusetts Institute of Technology, 2019

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2021

Author . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
Department of Electrical Engineering and Computer Science
August 6, 2021

Certified by . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
James R. Glass
Senior Research Scientist
Thesis Supervisor

Certified by . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
Yonatan Belinkov
Senior Lecturer
Thesis Supervisor

Accepted by . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
Katrina LaCurts
Chair, Master of Engineering Thesis Committee

# Constructing Low Resource Approaches to Improve Speech-to-text Translation from Modern Standard Arabic to English

by

Rami Manna

Submitted to the Department of Electrical Engineering and Computer Science
on August 6, 2021, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

## Abstract

This thesis explores novel approaches to the Arabic-English speech-to-text translation task. First, we construct a novel Modern Standard Arabic speech and English text parallel dataset. Second, we propose a novel framework for leveraging unsupervised machine translation to improve speech-to-text translation, and apply this framework to the task of Arabic-English speech-to-text translation. In particular, we propose a 3-step cascade approach to speech-to-text translation. In step 1, we use a speech recognition model to transcribe the Arabic speech into Arabic text. In step 2, we leverage unsupervised machine translation to learn a mapping between the output of the speech recognition model (transcribed Arabic) and Modern Standard Arabic (formal written Arabic). In step 3, we use an Arabic-English machine translation model to translate the output of the unsupervised model to English. Our third contribution is an exploration of approaches to low-resource end-to-end speech-to-text translation. We present and compare two approaches for synthesizing parallel training data. Finally, we compare the end-to-end approach with the cascaded approach. We found that the 3-step cascaded speech-to-text did not perform as well as the 2-step cascaded speech-to-text baseline. We show that with the end-to-end approach trained with synthetic English text, we are able to achieve similar performance to the 2-step cascaded speech-to-text baseline.

Thesis Supervisor: James R. Glass
Title: Senior Research Scientist

Thesis Supervisor: Yonatan Belinkov
Title: Senior Lecturer

# Acknowledgments

I would like to thank my advisors, **Jim Glass** and **Yonatan Belinkov**, without whom this thesis would not have been possible. Thank you for giving me the incredible opportunity to work with you, and for your fantastic mentorship. Additionally, I am grateful for my collaborator, Sameer Khurana, who I have learned a lot from over the course of my M.Eng. I am also grateful for the feedback and expertise of my labmates at the Spoken Language Systems group. Thank you to DSTA for supporting my appointment as a Research Assistant.

I would also like to give a special thank you to my family. Thank you, ميسون (Maysoun), أحمد (Ahmad), أمين (Amin), فادي (Fadi) and عزيزة (Aziza) for your limitless love and support. Last but not least, thank you to my incredible friends, Alyssa, Mayuri, Felipe, Sabine, Lisa and Amir for their moral support and invaluable feedback and advice throughout my M.Eng. I am lucky to have you all.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## 1.1 Context and Motivation

With the advent of internet globalization, international travel and collaboration, cross-lingual communication has become essential. The impacts of language barriers range from misunderstandings in conversations to differences in access to quality healthcare [1].

The vast majority of research in natural language processing (NLP) has studied the English language. As the official language of 22 countries, Arabic has over 400 million speakers worldwide [2], making it among the 6 most widely spoken languages. Work on Arabic-English translation has the potential to break down the language barrier between the Arabic speaking world and the English speaking world of over 1.1 billion speakers.

The task of translating from Arabic is one that has attracted a lot of attention from the scientific community due to the difficulty of the Arabic language. Together with the fundamental differences between the Arabic and English languages, this makes advances in computationally producing high quality Arabic-English translations imperative for extending the capabilities of computational linguistics.

Studying Arabic speech in particular, is interesting because of the amount of variation in how Arabic is spoken. Spread across 22 Arabic speaking countries, Arabic has a rich and diverse range of spoken dialects. Unlike the minor differences between spoken English in major English speaking countries, the large dialectal variation in Arabic means that only some of the Arabic dialects are mutually intelligible (e.g. Lebanese and Egyptian), while many are not (e.g. Moroccan and Palestinian).

In addition to colloquial Arabic dialects, which are informal in nature, Arabic also has a formal dialect, referred to as Modern Standard Arabic (MSA). MSA is used across the entire Arabic speaking world for written Arabic, news broadcasting, and formal communications. It is also an enabler of spoken communication between Arabic speakers of mutually unintelligible Arabic dialects.

Despite the richness of spoken Arabic, most Arabic NLP research has focused on Arabic text. Similarly, most machine translation research involving Arabic has focused on translation to and from Arabic text. Additionally, the focus has largely been on MSA, which is not a spoken language. This is largely because MSA has large amounts of data available to support research.

While Arabic Automatic Speech Recognition (ASR), the automatic transcription of Arabic speech into Arabic text, has seen substantial improvements in recent years, not much research has been done on translation from Arabic speech to other languages.

## 1.2 Arabic-English Speech-to-text Translation

In this thesis, we focus on translation from MSA speech to English text. Beyond the decision to focus on speech translation, this involves two significant decisions: 1) studying MSA instead of dialectal Arabic, and 2) focusing on translating from Arabic to English, as opposed to English to Arabic. We explain our rationale for making these decisions in the two subsections below.

### 1.2.1 Rationale for Studying Modern Standard Arabic

Dialectal Arabic is by far the dominant form of spoken Arabic. Therefore, our long-term goal is to architect speech translation systems that are capable of translating dialectal Arabic speech. However, the translation of dialectal Arabic speech is challenging because there is very little data to support research. In particular, since there is a large variety of Arabic dialects, and many of these dialects are mutually unintelligible, we would need a significant amount of data for each individual dialect. Most individual dialects have relatively a small number of speakers, and therefore have very little data available for research.

Due to the low availability of resources for dialectal Arabic speech translation, we instead focus our effort on MSA-English speech-to-text translation, for which enough data is available for research. However, MSA-English speech-to-text translation is still a low-resource task, and we devote considerable effort researching methods for overcoming data scarcity throughout this thesis.

### 1.2.2 Rationale for Arabic to English Translation Direction

We choose to investigate the Arabic to English direction of translation primarily due to the difficulty of translating in the reverse direction, from English to Arabic. This difficulty stems from the fact that Arabic is much more morphologically rich than English. For example, generating a single translated word might involve choosing both the correct stem and inflection. While translating to morphologically rich languages poses interesting challenges, we choose to focus on the challenges of speech-to-text translation instead.

## 1.3 Challenges of Translating Arabic

Translating Arabic gained attention early on as one of the major challenges for machine translation. Arabic differs from many other languages mainly due to its complex morphology and diacritization [3]. The rich set of differences between Arabic and English makes translating between Arabic and English particularly challenging. Below is a summary of the main differences between the two languages:

- Unlike English, Arabic is a Morphologically Rich Language (MRL) [4], which means that significant information about syntactic relations is often embedded at the word level (as opposed to the sentence level). For example, استقبلوه (gloss: istaqbaluhu) translates to "they welcomed him" in Arabic. This is a case where the subject, verb and object of the sentence are all parts of the same word.

- Because of the complex morphology, one word in Arabic often translates into multiple words in English. For example, أكلته (gloss: akaltuha) translates into three English words, "I ate it", while ضحكوا (gloss: dahaku) translates into two words, "they laughed". This makes translation more difficult because learning a one-to-one mapping between words no longer suffices. One effective approach for alleviating this difficulty is preprocessing Arabic text by dividing words into meaningful sub-word units.

- Affixation is found in both English and Arabic, but is substantially more complex in Arabic. Affixes in English can only involve the addition of the affix to a word stem. Affixation in Arabic can involve insertions, deletions or substitutions. Infixes, which involve insertions directly into a word stem, are common in Arabic but rare in English [5].

- Inflections are a lot more common in Arabic than in English [5]

- In Arabic, there is a large space of morphological forms for a given lemma. For example, “I walked”, “they walked”, “she walked”, and “he walked” are all different forms in Arabic, whereas the same form, “walked”, is used in English for all of these cases.

- Arabic has diacritics, which are small marks placed directly above or below individual characters. Their purpose is to describe what sound to make between the current and next character. The most common diacritics represent shorter versions of vowels. In addition to guiding pronunciation, these diacritics are sometimes responsible for resolving ambiguities. For example, a diacritic might signal the gender of a subject or object. These diacritics are often left out of Arabic text, creating syntactic and morphological ambiguities, that can make translation more difficult [3]. In practice, these ambiguities are resolved by Arabic readers based on context.

In addition to the general challenges of translating Arabic, speech-to-text translation introduces additional unique challenges relative to MT tasks. In particular, MT is text-based, and therefore deals with grammatical, punctuated text, and generally has large resources available. Meanwhile, speech is agrammatical, and does not contain punctuation. This mismatch between how language is structured in speech and text adds a layer of complexity for speech-to-text translation. In the case of Arabic, even more complexity is introduced because the common written form, MSA, is not spoken in general.

## 1.4 Thesis Contributions

Our first main contribution is the construction of a novel Modern Standard Arabic speech and English text parallel dataset. Our second contribution is a novel framework for leveraging unsupervised machine translation to improve Arabic-English speech-to-text translation. As our third contribution, we explore approaches to low-resource end-to-end speech-to-text translation via synthesizing parallel data. We find that synthesizing English text from speech recognition datasets to create pseudo-parallel data works best, achieving similar translation quality to that of the 2-step cascaded ST approach. We also explore an extension of this end-to-end approach by filtering the synthetic pseudo-parallel data by quality.

## 1.5 Thesis Overview

In Chapter 2, we give a background of related work. In Chapter 3, we describe a novel Arabic-English speech-to-text translation dataset. In Chapter 4, we propose a 3-step approach to cascaded speech-to-text translation. In Chapter 5, we explore approaches to low-resource end-to-end speech-to-text translation. In Chapter 6, we summarize our contributions and point to future work.

# Chapter 2

# Background

In this chapter, we give a background on the state-of-the-art methods for machine translation using neural networks. We discuss the challenges of machine translation and speech-to-text translation that arise when translating from Arabic to English in particular.

## 2.1 Machine Translation

Machine Translation (MT) involves writing computer programs that are able to translate sentences from a source language to a target language. MT research has predominantly focused on settings where both the source and target are natural languages. Additionally, most MT research has focused on translating text, as opposed to other language mediums such as speech. For example, an MT system might attempt to translate an English text sentence into French text.

### 2.1.1 Neural Machine Translation

In recent years, the field of machine translation has seen large breakthroughs owing to successes in the use of neural networks for natural language processing (NLP). In the case of machine translation, these neural machine translation (NMT) models are sequence-to-sequence models; that is, their input and output are both sequences.

These models are generally supervised, meaning that they depend on labeled data to be trained. In particular, the models require training data in the form of pairs of parallel sentences: a sentence in the source language, and its corresponding translation in the target language. Generally, models with more training data produce better translations. The amounts of parallel training data available may vary for different language pairs, making training a translation model for lower-resource language pairs more difficult than for those with plentiful training data available.

Until recently, recurrent neural network (RNN) architectures such as the Long-Short Term Memory (LSTM) architecture, were the state-of-the-art sequence-to-sequence models for most NLP tasks. However, following the introduction of the attention mechanism [6], the transformer architecture [7] has largely replaced recurrent models as the state-of-the-art across NLP tasks, including MT. The transformer architecture is composed of an encoder, which translates the source sentence into an intermediate representation, and a decoder, which produces the translated sentence form the intermediate representation. One of the main distinguishing features of the transformer model is that it uses an attention mechanism to allow the decoder to selectively attend to the parts of the encoder output that are most useful for generating its translations.

### 2.1.2 Consequences for Speech-to-text Translation

Neural networks in the context of speech recognition systems were first introduced in the 1990s [8, 9, 10]. The fact that they have now been shown to also drastically improve machine translation is quite profound, especially in the context of translating speech from one language into text in another, as it implies that there is a learnable underlying structure between languages.

Furthermore, it is interesting that the same computational model, the neural network, is able to achieve state-of-the-art performance on language tasks involving both speech and text representations of language. This shows potential for neural networks to be used for developing a single end-to-end neural model that spans both text and speech representations of language.

## 2.2 Arabic-English Machine Translation

### 2.2.1 MSA vs. Dialectal Arabic Research

The majority of work on Arabic-English machine translation has been text-based, and therefore focused on MSA [11]. This is because MSA is the primary form of written Arabic, whereas dialectal Arabic is almost exclusively spoken, and rarely written [12], which makes research on dialectal Arabic MT difficult because of the lack of written dialectal Arabic text corpora. With increased internet usage, written dialectal Arabic has become more common in recent years, enabling some research in dialectal Arabic MT, but the majority of research and data available is still in MSA.

Modern transformer-based approaches to Arabic-English MT achieve Bilingual Evaluation Understudy (BLEU) scores, the currently prevalent metric, of 31.1 [13].

### 2.2.2 Word Segmentation

Word segmentation is the process of segmenting words into smaller parts, or "subwords". Since Arabic is morphologically rich, word segmentation plays a critical role in Arabic NLP. As a result, a lot of emphasis on developing morphological segmentation tools such as MADAMIRA [14] and Farasa [15]. These segmenters are specifically designed for MSA, and have been shown to perform significantly less well on dialectal Arabic [16]. Recently, Sajjad et al. showed that using a Byte-Pair Encoding (BPE) [17] for Arabic-English MT results in similar performance to state-of-the-art language-dependent segmenters [13].

### 2.2.3 Dataset Availability

Sajjad et al. aggregated MSA-English parallel datasets, resulting in a combined 41 million parallel sentences [13]. This is much more parallel data relative to that available for individual Arabic dialects.

## 2.3 Arabic-English Speech-to-text Translation

In contrast, Arabic speech translation has been scarcely researched. We are aware of only a few such efforts. The first is a submission to the Arabic-English speech-to-text translation task of the IWSLT (International Conference on Spoken Language Translation) 2007 challenge [18, 19]. The authors use a Statistical Machine Translation (SMT) approach, and re-rank their translations with additional models. The second is a study of Iraqi-Arabic to English speech translation [20]. This is different from our setting because it focuses on dialectal Arabic speech rather than MSA speech. As these papers are over a decade old, we were unable to directly compare our work to theirs. Both papers use SMT, an approach that has since become outdated due to the successes of neural approaches. We are not aware of any follow-up Arabic-English speech-to-text translation research using modern (neural) approaches.

### 2.3.1 Dataset Availability

The largest Arabic-English speech-to-text translation dataset we are aware of is the Egyptian CALLHOME dataset [21]. It is comprised of pairs of Egyptian Arabic Speech utterances and their corresponding translations into English text. The training data comprises of 20 thousand utterances, and the development and test sets make up another 15 thousand utterances. Importantly, the speech in this dataset is dialectal Arabic speech, which differs significantly from MSA speech.

More recent work in 2020 introduces an MSA-English speech-to-text translation dataset as part of CoVoST 2, a massively multilingual speech-to-text translation dataset [22]. This MSA-English dataset is made up of a 2283 sentence training set, 1758 sentence development set and a 1695 sentence test set. The paper reports BLEU scores evaluated on their test set ranging from 0.1 with a cascaded ST approach, to 4.3 with an end-to-end ST approach. Clearly, the size of the CoVoST 2 training dataset is too small for training a quality speech-to-text translation system.

We are not aware of other Arabic-English speech-to-text datasets. The datasets available are not sufficient to support research in MSA-English speech-to-text translation. The lack of data is a major challenge that we seek to overcome throughout this thesis, both by collecting a new dataset and exploring techniques for perfoming low-resource translation.

## 2.4 Evaluation Metrics

A key enabler of conducting research in speech-to-text translation is having the ability to quantitatively evaluate a translation model in an automatic fashion. This allows researchers to iteratively improve models without the need for continuous manual inspection of the translations the model produces. Having a standardized automatic metric for measuring translation quality also enables different researchers to efficiently collaborate and compare their models in an objective manner.

The current prevalent metric for automatically evaluating machine-generated translations is the Bilingual Evaluation Understudy (BLEU) metric [23]. The BLEU score for a given machine-generated translation of a source sentence is calculated based on its similarity to one or more human-generated translations of the same source sentence. Machine-generated translations that are similar to the human-generated translations receive higher BLEU scores. This similarity is computed based on the number of n-grams (sequences of consecutive words) that the machine and human generated translations share. BLEU scores range from 0 to 100, with higher scores representing more accurate translations.

Notably, the calculation of a BLEU score depends on having human-generated translations for use as a reference. The parallel datasets discussed in the Dataset Availability sections of this chapter are composed of source sentences and their corresponding human-generated reference translations. As we saw in this section, in addition to their use for training models, datasets are also needed for evaluating trained models.

## 2.5 Chapter Summary

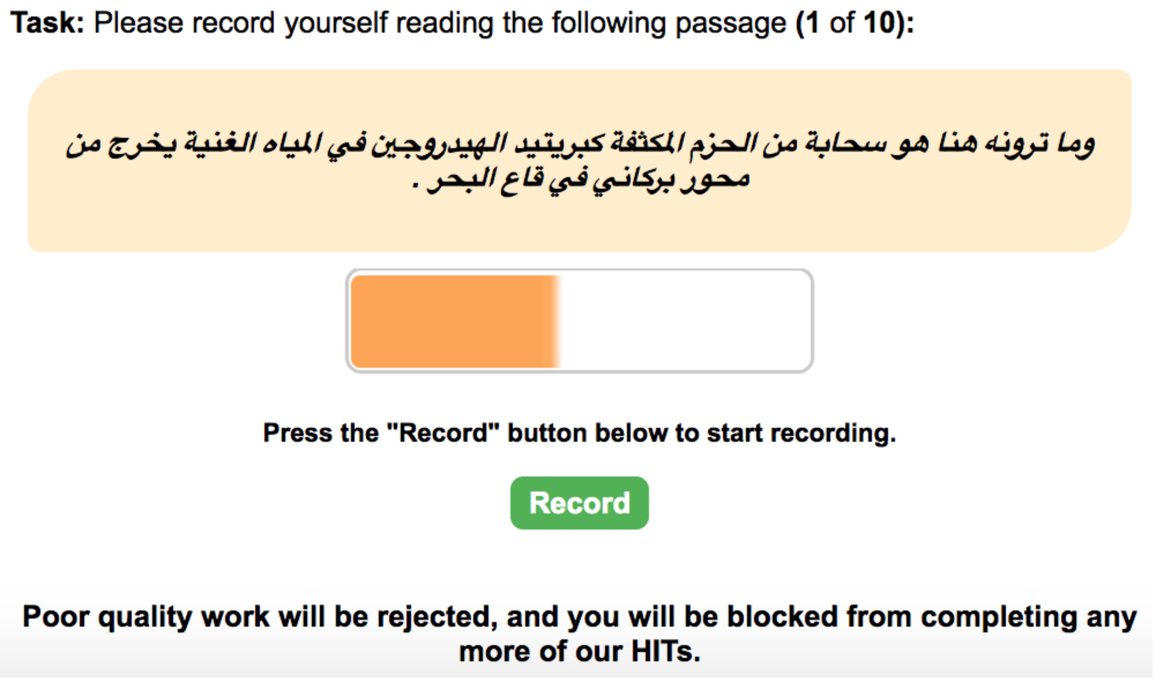
In this chapter, we discussed the current state-of-the-art for MT, which involves the transformer neural architecture, and its use of attention. We further investigated Arabic-English machine translation and speech-to-text translation, for which we discussed the choice of Arabic dialect, evaluation metrics, word segmentation techniques and the availability of relevant datasets to support research.

# Chapter 3

# Novel MSA to English Speech Translation Dataset

## 3.1 Motivation

In order to evaluate our speech-to-text translation system, we needed a test set in the form of parallel pairs of MSA speech utterances and their English text translations. To the best of our knowledge, no such corpus of significant size exists. The closest contender to such a corpus is the "The CALLHOME Egyptian Arabic Speech Translation Corpus" [21], which is comprised of pairs of Egyptian Arabic speech utterances and their corresponding translations into English text. The dataset is made up of 35,842 utterances, and is partitioned into training, development and test sets. While the development and test sets of this dataset are sufficiently large for evaluating speech-to-text translation system, the dataset is not a good fit for our use case because the Egyptian dialect differs substantially from MSA. Therefore, it is unlikely to work well for evaluating an MSA to English speech-to-text translation system. With no dataset consisting of MSA speech utterances and their corresponding English text translations, it is difficult to make progress on the end-to-end approach, therefore we decided to collect a test dataset.



**Task:** Please record yourself reading the following passage **(1 of 10):**

وما ترونه هنا هو سحابة من الحزم المكثفة كبريتيد الهيدروجين في المياه الغنية يخرج من
محور بركاني في قاع البحر .

**Press the "Record" button below to start recording.**

**Poor quality work will be rejected, and you will be blocked from completing any more of our HITs.**

Figure 3-1: Example of Data Collection Prompt

## 3.2 Dataset Construction

To collect the dataset, we launched a task on Amazon Mechanical Turk (AMT) in which we asked crowd source workers to record themselves reading Arabic sentences out loud. Figure 3-1 shows the user-interface presented to crowd source workers, with an example sentence prompt to be recorded. We chose these Arabic sentences from our 41 million sentence parallel Arabic-English text dataset, which means that we automatically have the English translation for each Arabic speech utterance we collect. Another benefit of this data collection setup is that we are not only getting Arabic speech to English text parallel pairs, but we also have their corresponding ground truth Arabic text.

### 3.2.1 Dataset Size

We collected a total of 26,633 utterances. We collected 19941 of these utterances as part of a train dataset, and the remaining 6692 utterances as a test dataset.

### 3.2.2 Sentence Length

In choosing Arabic sentences to use for the speech data collection task, we limited our choices to sentences between 5-20 words. This is mainly because reading a sentence longer than twenty words is an arduous task, and we want to make sure the AMT task remains manageable.

### 3.2.3 Data Diversity - Worker Demographics

MSA is spoken throughout the Arabic-speaking world, so in order to make sure our test dataset is representative of MSA, we opened up the task to workers from all of the countries in the Arab world, as well as the US because the vast majority of AMT workers reside in the US [24]. We have recorded utterances from Jordan, Lebanon, the United States, Bahrain and Palestine. We initially launched the task with the description and instructions in English, which may have been limiting the number and/or diversity of workers that we were receiving. Therefore, in order to increase the volume and diversity of workers, we internationalized the data collection website to have both a version with an English description and a version with an Arabic description.

### 3.2.4 Data Diversity - Audio and Equipment

The speech utterances are recorded through a web interface through the microphones on the various personal computers of the AMT workers. The recordings are therefore recorded with a diverse set of computer microphones and in various levels and types of background noise. However, we do ask the workers to record in a relatively quiet place so the surrounding noise is minimal in general.

### 3.2.5 Limitations

The main caveat of the dataset we collected is that the Arabic speech utterances are read as opposed to naturally spoken. It is possible that the distribution of "read" speech differs from that of naturally spoken speech. For example, repetition of a sentence or visual reading errors may be more prevalent in read speech. On the other hand, disfluencies may be less common in read speech because a well-formed grammatical sentence is already provided to the speaker.

## 3.3 Quality Assurance

Since there is a possibility that some crowd source workers may not read out sentences properly, we need a way to ensure that the utterances included in the dataset are high quality so that they can serve as a authoritative ground truth for evaluating our speech-to-text translation system. Therefore we take the following steps to ensure the accuracy of crowd sourced utterances:

- **Length Threshold:** Recordings less than 1 second long are discarded.

- **Real-time Validation:** Utterances are transcribed using the Google Speech-to-Text API, and the word error rates (WER) of the transcriptions are calculated. Only utterances whose transcripts result in a WER of less than 40% are kept. We chose this threshold manually by inspecting the quality of the speech utterances produced. The main goal of this threshold was to ensure that crowd source workers were in fact reading the sentence presented and not trying to game the system.

- **Task Approval Rating:** AMT allows the assignment of tasks to be restricted to crowd source workers that meet a certain approval rating, i.e. a score representing the percentage of tasks they have performed satisfactorily in the past. When we started the data collection, we only allowed workers with an approval rating of 90% or greater to attempt our task. As there are not many crowd source workers from Arabic speaking countries, this high approval rating threshold led to a very low rate of data collection. To expedite the data collection process, we lowered the approval rating threshold to 65% and rely more on real-time validation of incoming speech utterances to maintain quality.

## 3.4 Chapter Summary

In this chapter, we described the collection of a novel MSA-Arabic speech-to-text translation dataset. We detailed the dataset construction process, highlighting the steps we took towards data diversity and quality assurance. The training set we collected contains 40.45 hours of speech, and the test set contains an additional 12.96 hours, resulting in a total of 53.41 hours across the entire dataset. Table 3.1 summarizes the key details of the dataset.

| Dataset | # Utterances | Ave. Words/Utterance | Ave. Duration (seconds) |
|---|---|---|---|
| Train | 19,941 | 11.66 | 7.30 |
| Test | 6,692 | 11.27 | 6.97 |

Table 3.1: Summary of novel Arabic-English speech-to-text translation dataset

# Chapter 4

# 3-step Cascade Speech-to-text Translation

In this chapter, we explore a novel approach to performing cascaded speech-to-text (ST), in which the 2-step cascade ST is augmented with an intermediate step responsible for aligning the output of the ASR step with the input of the MT step.

## 4.1 Baseline: 2-step Cascade Speech-to-text Translation

In the absence of sufficient parallel data for training an end-to-end system, the dominant approach to speech-to-text translation is cascaded ST. Speech-to-text translation systems are often designed as 2-step cascades. Cascaded ST is a system made up of two consecutive models: an ASR model followed by an MT model . We introduce a novel Arabic-English 2-step cascade model as our baseline. We construct the system as follows:

### 4.1.1 Automatic Speech Recognition Model

We use an in-house Arabic ASR model developed by Sameer Khurana. The model is trained on the MGB-2 speech recognition dataset using the ESPnet framework.

Specifically, we use the “Speech-transformer” ASR architecture [25, 26], and train the model with a hybrid CTC/attention loss function [27]. The architecture consists of a convolutional pre-encoder, a transformer encoder and transformer decoder. The encoder has 12 layers and the decoder has 6 layers, each with 2048 hidden units. Overall, the model has approximately 200 million learnable parameters. The model is trained on 1200 hours of Arabic speech from the MGB2 dataset, and achieves a WER of 12.5% on the MGB2 test set.

### 4.1.2 Machine Translation Model

We use the Arabic-English MT model described by Sajjad et al. [13]. It is a transformer-based model trained on the concatenated dataset of 41 million Arabic-English sentence pairs. The trained model achieves an average BLEU score of 31.1 on four IWSLT test datasets.

## 4.2 Motivation

While both are forms of Arabic text, transcribed Arabic speech and written MSA are different in many ways. For example, transcribed Arabic speech does not contain punctuation, may have repeated words and is more likely to have grammatical mistakes or partial sentences. Written Arabic, on the other hand, tends to be more structured, comprising of full and grammatically correct sentences. We can mitigate some of these differences by choosing a formal Arabic speech dataset in which the Arabic speech was the result of an Arabic speaker reading written Arabic. We refer to this as “read speech”. In particular, we could use audio from Arabic news broadcast in which the news anchor reads written Arabic off of a teleprompter. Alternatively, we could use Arabic audiobooks, in which a narrator reads written Arabic text from a book. However, significant differences between the Arabic speech transcriptions and Arabic written text are inevitable. Such differences could be due to the lack of punctuation in the transcribed Arabic, any mispronunciations or mistakes made by news anchors, or transcription errors made by the ASR system.

Hence, there is a misalignment between the distributions of the output of the ASR model (transcribed Arabic) and the input the MT model expects (MSA text). However, the 2-step cascaded ST system is does not take this misalignment into account, as the transcribed Arabic output of the ASR model is fed directly as input for the MT model even though it is not MSA. We hypothesize that reducing this misalignment should improve the performance of the MT model, and therefore result in an improvement in the overall system's speech-to-text translation quality.

## 4.3 Oracle Experiment

We start by running an oracle experiment to evaluate the potential of correcting the misalignment between the ASR and MT steps of the cascade ST. To quantify this potential, we simulate what would happen if we were able to perfectly fix the misalignment issue. That is, we assume that we were able to perfectly convert the ASR output into MSA. To achieve this, rather than feeding the imperfect ASR output into the MT model, we instead feed the ground truth MSA into the MT model. Finally, we evaluate the BLEU score on our test set. This oracle experiment results in a BLEU score of 27.5. This shows that improving the alignment of the ASR output with the MT input has a potential for an improvement of up to 12.8 BLEU points from the baseline of 14.7 BLEU. This large potential for improvement validates that the misalignment problem is a significant one that is well worth studying.

## 4.4 Method

In order to reduce the misalignment, we propose to augment the cascaded ST approach with an intermediate step between the ASR and MT steps of the cascaded ST approach. The purpose of this intermediate step is to convert the transcribed Arabic output of the ASR, into MSA, the written style of Arabic the MT system is trained on. Our hypothesis is that this resulting improved alignment between the MT input and the distribution of the MT training data should improve the quality of the MT's final English translations.

We can think of this “style transfer” as a translation between two styles of Arabic (spoken style Arabic to MSA style Arabic). If we think of the task as a translation, we could even employ a machine translation algorithm to complete the task.

However, supervised machine translation algorithms are not an option because we do not have parallel data for this pair of Arabic styles of writing. We can create parallel transcribed Arabic to MSA training data from the training set of the speech-to-text translation dataset we collected. In particular, we constructed the dataset by tasking crowd source workers to read Arabic sentences from a parallel Arabic-English dataset. This means that for each Arabic speech to English text pair we collected, we also have the corresponding MSA sentence, resulting in triplets of Arabic speech, MSA text and English text. We transcribe the Arabic speech using our trained ASR model to generate transcribed Arabic sentences that correspond to the MSA text. Since the training set of the speech-to-text translation dataset we collected consisted of 19,941 utterances, this gives us 19,941 parallel pairs of transcribed Arabic to MSA text. However, 19,941 parallel sentences is not sufficient for training a supervised MT model. Therefore, we explore means of performing the translation without parallel data (unsupervised MT), or with limited parallel data (semi-supervised MT).

Fortunately, due to recent advances in unsupervised machine translation [28, 29, 30, 31], it has become possible to learn translations between two languages in the absence of parallel data. To compensate for the parallel data, unsupervised MT does require large monolingual corpora with 2.9 million sentences or more per language [30].

Lample et al. 2019 [31] achieve a BLEU score of 34.3 on the German-English WMT’16 translation task, establishing a new state-of-the-art for unsupervised MT. This result is particularly significant because this was a 9 BLEU point increase relative to the previous state-of-the-art unsupervised MT model. We use the unsupervised machine translation model they describe to “translate” the output of the ASR system into the input the of the MT system. The resulting augmented 3-step cascade consists of the following:

1. **Transcription:** Use an Automatic Speech Recognition system to transcribe the Arabic speech waveforms into Arabic text.

2. **Unsupervised Machine Translation:** Translate from transcribed Arabic into Modern Standard Arabic text using unsupervised MT.

3. **Translation:** Use an Arabic text to English text Machine Translation system to convert the Modern Standard Arabic output from step 2 into English text.

A visualization of this pipeline is rendered in Figure 4-1, and a visual comparison of the considered approaches is rendered in Figure 4-2.

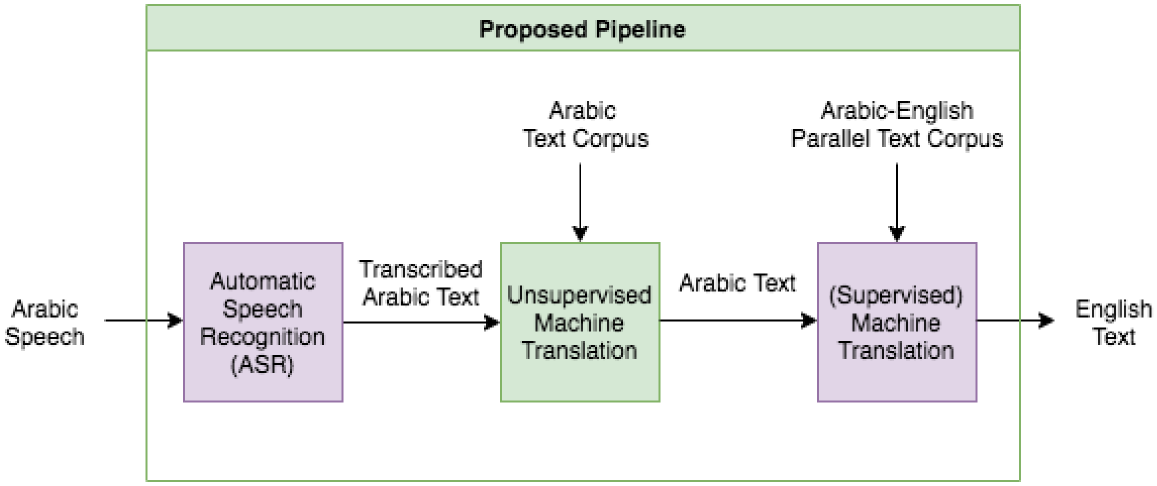

Proposed Pipeline
Arabic Text Corpus
Arabic-English Parallel Text Corpus
Arabic Speech
Automatic Speech Recognition (ASR)
Transcribed Arabic Text
Unsupervised Machine Translation
Arabic Text
(Supervised) Machine Translation
English Text

Figure 4-1: 3-step Cascaded ST Overview

## 4.5 Challenges

Monolingual data is much more common than parallel data, and so for many languages, sufficiently large monolingual corpora are available. This is the case for MSA, for which we have a large 41 million sentence MSA dataset. Since transcribed MSA speech is not a conventional language, no corpora of transcribed Arabic text exist. However, we can create a monolingual corpus of Arabic speech by using an ASR model to transcribe MSA speech. The amount of transcribed MSA text we can acquire in

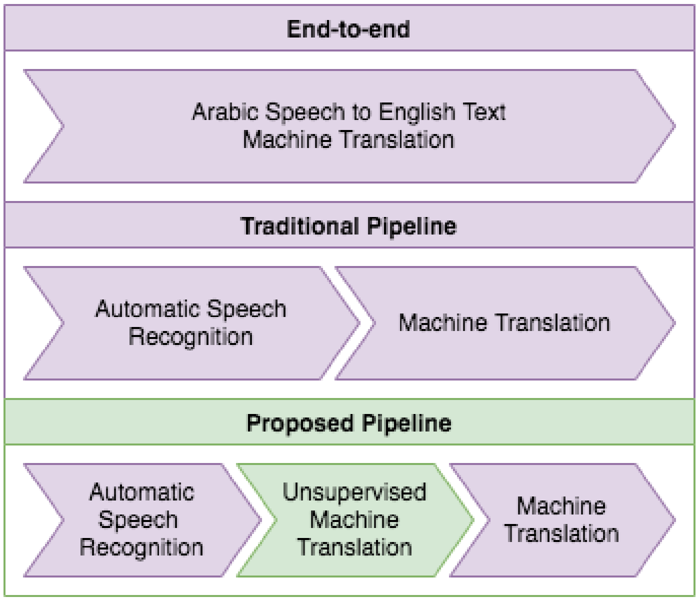

End-to-end
Arabic Speech to English Text Machine Translation
Traditional Pipeline
Automatic Speech Recognition
Machine Translation
Proposed Pipeline
Automatic Speech Recognition
Unsupervised Machine Translation
Machine Translation

Figure 4-2: Summary of Considered Approaches

this way is limited by the amount of MSA speech data we have. By transcribing the speech from the MGB-2 dataset, we create 375,103 sentences of transcribed MSA.

It is important to note that we are not using unsupervised machine translation in the typical way. While unsupervised machine translation was created for translating between completely different languages, such as English and French, our system translates between closely related forms of Arabic (transcribed Arabic and MSA). Unsupervised machine translation depends on the assumption that some commonality can be found between the languages that at some level they are structurally similar, so that the learnings from each corpus can be mapped onto each other. We expect transcribed Arabic and MSA to be more similar than, for example, English and French. Therefore, despite the relatively low availability of monolingual data for transcribed Arabic, we are optimistic about applying this approach because we expect the actual translation task to be easier, and thus require less data.

Another possible challenge might be the increased presence of noise or inaccuracies in the transcribed Arabic. In particular, erroneous transcriptions may be phonetically similar but semantically unrelated to the their source utterances. As a result, corpora of transcribed Arabic may conform less rigidly to expected linguistic rules.

## 4.6 Implementation

### 4.6.1 Preprocessing

Once we have gathered the datasets, the first step is to preprocess the data. Due to the morphological properties of Arabic, word segmentation is essential for Arabic NLP [32, 33] The most significant component of our preprocessing pipeline is segmenting words into subword units. We do this using the approach introduced by Sennrich et al. [17], which performs a word segmentation algorithm based on the byte pair encoding (BPE) compression algorithm. Since this method has its roots in information theory, it is completely language agnostic. Applying BPE has been shown to improve performance on Arabic English machine translation [13]. Therefore, we expect that it will work well for transcribed Arabic to MSA machine translation.

Figure 4-3 shows the steps we use for preprocessing. We start by tokenizing each of the Arabic text corpus and the transcribed Arabic text corpus. We then learn BPE over the concatenation of the Arabic text and transcribed Arabic corpora, and use those codes to apply BPE onto the tokenized version of each corpus. This produces tokenized and BPE encoded versions of the corpora, ready to be used as inputs for the Unsupervised Machine Translation step.

### 4.6.2 Adapting Unsupervised Machine Translation for Transcribed Arabic to MSA translation

While we have existing systems for the first and third steps of our pipeline (ASR and MT respectively), we need to build a system for step 2 of the pipeline: translating between transcribed Arabic and Modern Standard Arabic. We base our work on [31], which achieves the current state of the art for Unsupervised MT. In particular, we use Facebook Research's XLM code repository (https://github.com/facebookresearch/XLM) as a starting point.

XLM was designed for languages other than Arabic. This introduces more complexity because Arabic differs significantly from these languages, particularly due to

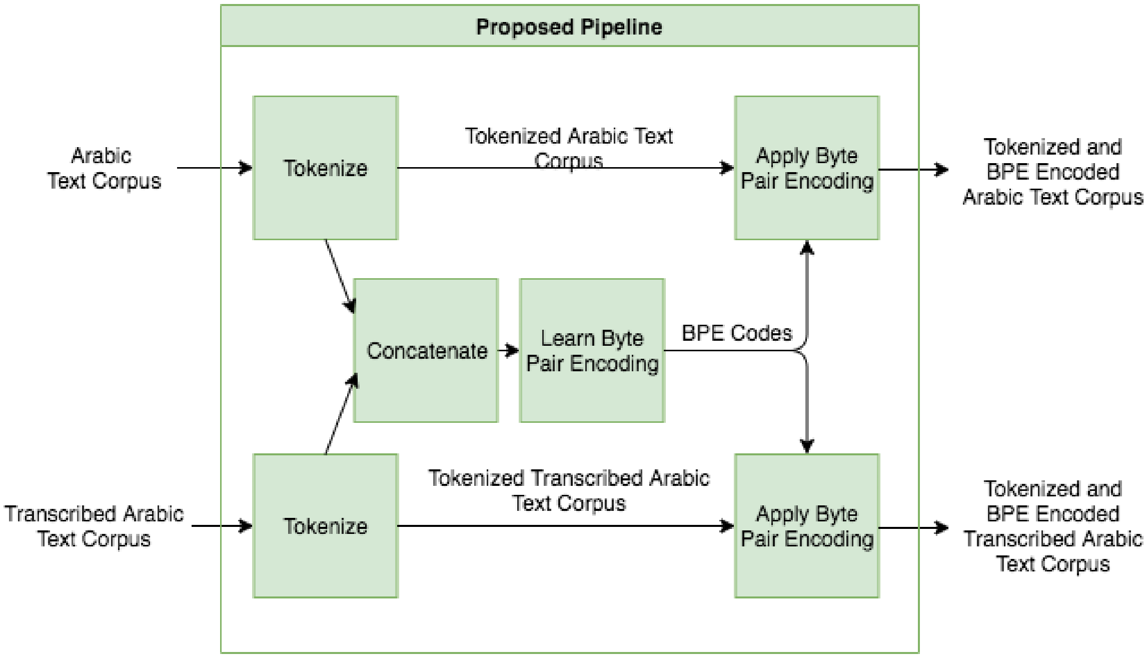

Proposed Pipeline
Arabic Text Corpus
Tokenize
Tokenized Arabic Text Corpus
Apply Byte Pair Encoding
Tokenized and BPE Encoded Arabic Text Corpus
Concatenate
Learn Byte Pair Encoding
BPE Codes
Transcribed Arabic Text Corpus
Tokenize
Tokenized Transcribed Arabic Text Corpus
Apply Byte Pair Encoding
Tokenized and BPE Encoded Transcribed Arabic Text Corpus

Figure 4-3: Preprocessing Overview

its more complex morphology. Therefore, it may be beneficial to use Arabic specific preprocessing methods to preprocess the data in a way that we know works well for Arabic text instead of using the preprocessing methods used by default by XLM. With our preprocessing methods integrated into the XLM code, we then tune the hyperparameters of the model for our language pair. In particular, since we are dealing with a translation between two styles of Arabic (as opposed to two completely different languages), the default hyperparameters are unlikely to work for our task.

## 4.7 Experiments

The first thing we do is test the proposed 3-step cascaded ST approach in its most basic form, to get baseline results and see if the approach is viable.

The approach is composed of 3 consecutive steps: Arabic ASR, Arabic Transcription to MSA Unsupervised MT, and Arabic-English MT. Since we already trained an ASR and MT model for the 2-step cascaded ST, we only have to train the intermediate unsupervised MT step to complete the cascade.

### 4.7.1 Data

To train the unsupervised MT model, we need a monolingual corpus of Arabic ASR transcriptions and a monolingual corpus of MSA.

Since Arabic ASR transcriptions are not a language in the conventional sense, no public corpus of Arabic ASR transcriptions exists. However, since Arabic speech corpora are available and we have a trained Arabic ASR model, we can create a corpus of Arabic ASR transcriptions by decoding Arabic speech with the ASR model. We decode the MGB-2 Al Jazeerah news broadcast speech dataset using the trained ASR model to obtain 375,103 sentences of Arabic Transcriptions. For later experiments, we transcribe an additional 308,576 utterances from the GALE dataset to create a combined Arabic ASR Transcription dataset of 683,681 sentences.

For MSA, we use the 41 million MSA sentences from the combined Arabic-English parallel text dataset.

Table 4.1 summarizes the datasets used and their respective sizes.

| Dataset | Size |
|---|---|
| Arabic-English Parallel Text | 41,425,346 |
| MGB-2 Arabic Speech Dataset | 375,103 |
| GALE Arabic Speech Dataset | 308,576 |

Table 4.1: Datasets used for 3-step cascaded-ST

### 4.7.2 Evaluation

Since Arabic ASR transcription is not a conventional language, there wasn't an existing parallel dataset of Arabic ASR transcription and MSA sentence pairs that we can use to quantitatively evaluate the unsupervised MT model. However, we can create such a dataset by adapting the test set of the newly collected dataset described in Chapter 3. We do this by transcribing the read speech with the trained ASR model to produce a test set of parallel sentence pairs of Arabic ASR transcript - MSA sentences.

With a test set in hand, we evaluate the unsupervised MT model by translating the Arabic ASR transcripts from the test set and comparing the resulting output to the corresponding ground truth MSA to computing a BLEU score.

In comparing the ASR output against the ground truth MSA, we compute a score of 31.35 BLEU. This serves as a baseline for the intermediate step. That is, if the intermediate step did nothing, this is how similar the ASR output would be to MSA. This baseline means that the BLEU score of the intermediate step against the ground truth MSA needs to be greater than 31.35 for the intermediate step to have been successful in increasing the similarity to MSA.

After running many experiments and tuning hyperparameters, the best BLEU score we are able to achieve for the intermediate step in a fully unsupervised setting is 31.99. This increase in 0.64 BLEU points over the baseline indicates that the intermediate step has been at least partially successful in processing its input to become more similar to MSA. While evaluating the intermediate step in isolation is useful during development, the real measure of success is whether the addition of the intermediate step to the 2-step cascaded ST to form the 3-step cascaded ST results in an overall improvement in the Arabic-English speech-to-text translation. To do this, we again use our collected test set, and run the Arabic speech through the cascaded ST, once without the intermediate step, and once with the intermediate step. For each of these runs, we measure the BLEU score of the overall speech-to-text translation into English. The 2-step cascade baseline gives a BLEU score of 14.7, while the 3-step cascade results in a BLEU score of 14.3. That is, the small 0.54 BLEU improvement in the intermediate step's BLEU did not translate to a better overall speech-to-text translation system. Since 0.54 BLEU is quite a small change, it is possible that the intermediate step is not making meaningful changes, but is simultaneously making other changes that corrupt the sentence. Since BLEU operates on a word level and is agnostic of actual words, it could be that the intermediate step makes a larger number of improvements in less semantically significant words, such as punctuation or stop words, and a relatively smaller number of corruptions of more semantically (or syntactic) significant words such as nouns and verbs. This would result in a small increase in BLEU score even though the sentences have largely lost semantic content or coherence.

Since we do have 19,941 sentences of training data, we also train a semi-supervised MT model that relies on the same monolingual corpora as the unsupervised model, but also makes use of parallel training data.

## 4.8 Analysis

The baseline experiments in the previous section showed that while the intermediate step can lead to a 0.54 point improvement in the BLEU score of the Arabic ASR transcription computed against the ground truth MSA, the overall speech-to-text translation BLEU score does not improve. We hypothesize that mistakes made in the intermediate step's translations are more linguistically significant than the successful corrections, which results in a misleading small increase in BLEU score instead of a decrease.

Additionally, we expected the semi-supervised model should perform better than an unsupervised model because it has access to a direct mapping between sentences of the two languages. However, with 25,000 parallel sentence pairs, the semi-supervised MT falls short of the unsupervised MT with a BLEU score of 31.49. This is possibly due to the parallel dataset of 25,000 not being large enough. However, we do not know for certain why the semi-supervised model underperforms the fully unsupervised model.

A major limitation of our system is that the size of our monolingual Arabic ASR transcription dataset is much smaller than the size of our monolingual MSA dataset, and an order of magnitude smaller than the sizes of the monolingual datasets used in the literature [29].

## 4.9 Synthetic Parallel Data Augmentation Experiments

To make up for the limitations in dataset size outlined in the previous section, we build on the baseline semi-supervised model of the previous section via parallel data augmentations. In general, the purpose of these data augmentations is to simulate pairs of parallel sentences where in each pair, one sentence resembles Arabic ASR transcriptions, and the other sentence resembles MSA. We explore 3 approaches to these parallel data augmentations and compare their results.

In the previous section, one of the major limitations identified was that we lack a sufficient amount of parallel data for our task. Collecting large amounts of quality Arabic speech data is costly and time consuming, so we proposed to address this limitation via synthetic data augmentation. To address the lack of parallel sentences, we analyze the ASR output data and identify opportunities for synthetic data creation. We define and compare several data augmentation / synthetic parallel data creation techniques:

### 4.9.1 Identity Parallel Data Augmentation

Our setting differs substantially from what is described in unsupervised translation literature, in the sense that the source and target are both variants of the same language in our case. Thus, we augmented the model with parallel sentences mapping to themselves. While some of these showed minor increases in BLEU scores for the unsupervised MT step, the improvements did not carry over to the overall pipeline.

To identify why unsupervised MT was not performing well for our use case, we compare our setting with the successful setting used in the unsupervised MT literature and identify two key differences: 1. The original paper used much larger monolingual datasets 2. The original paper focused on translating between two completely different languages, whereas our goal is to translate between two forms of Arabic. Specifically, the unsupervised MT model as described in the original paper does not have a prior that its source and target languages are variants of the same language, and is therefore not learning to copy sentences that were already well formed. To introduce this source-target language similarity prior, we augment the unsupervised MT model with parallel sentences mapping MSA sentences to themselves.

### 4.9.2 Punctuation-based Parallel Data Augmentation

Since Arabic ASR transcription does not contain any punctuation, we create synthetic parallel sentences by stripping the punctuation from MSA sentences to form corresponding synthetic ASR transcriptions.

### 4.9.3 Corruption-based Parallel Data Augmentation

To understand the nature of the parallel data required for our task, we analyzed the ASR output and ground-truth Arabic and quantified the differences between the ASR output and MSA. We identified two key patterns:

1. Erroneous words are often off by one or two characters and occur in consecutive chunks
2. Character differences are often phonologically similar
3. Words are often repeated
4. Consecutive sequences of words are often dropped, and such drops are more common at the beginning or end of a sentence

Based off of these observations, we designed corruption based synthetic parallel data augmentation techniques. To synthesize these parallel data, we start with MSA sentences from our Arabic-English parallel text dataset, and corrupt them in ways that correspond to the above observations, and then use the corrupted sentence and original MSA sentence as a parallel pair. For example, for the observation that sequences of words are often dropped, we might delete the first few words of an MSA sentence to form the corresponding synthetic ASR Arabic transcription sentence.

As a second example, based on the observation of single character differences between words in the Arabic ASR transcription and MSA, and leveraging our knowledge of Arabic — that its orthography and phonology are similar, we employ edit-distance as a metric for corruption. Specifically, we synthesize parallel data by starting with MSA sentences from our Arabic-English parallel text dataset, and randomly select and replace a small number of words in each sentence with a word that has an edit distance of 1 or 2 from the original word. This creates the corrupted version of a given sentence, which represents the ASR Arabic transcription, which is paired with the original MSA sentence to create a parallel pair. Upon further inspection, we noticed that there are few key differences between Arabic orthography and phonology, and decided to modify our approach by using a phonologically-weighted edit distance to account for these differences.

### 4.9.4 Results

Among the synthetic parallel data augmentations, the best intermediate MSA BLEU score was 33.50 with the 1 million synthetic parallel sentence pairs, yielding an improvement of 1.51 BLEU points over the best unsupervised model for the conversion to MSA style text. However, this improvement in the intermediate step did not translate to a better overall speech-to-text translation quality, at a BLEU score of 14.1.

## 4.10 Chapter Summary

In this chapter, we proposed a novel approach to augmenting cascaded ST with an unsupervised MT intermediate step to align the output of the ASR model with the input of the MT model. Additionally, we explored novel approaches of augmenting the intermediate-step model with synthetic parallel data and improved its BLEU score by 1.5 points. Overall, we found that despite the increase in BLEU score of the intermediate-step, the 2-step cascaded ST still performed best.

Table 4.2 summarizes the results for all experiments in this chapter.

| Experiment Description | ASR output to Arabic BLEU | English BLEU |
|---|---|---|
| Baseline: 2 step pipeline | 31.35 | 14.7 |
| Oracle: Perfect XLM step | 100 | 27.5 |
| Unsupervised XLM | 31.99 | 14.3 |
| Semi-Supervised XLM | 31.49 | 14.1 |
| Synth Para Data XLM | 33.5 | 14.1 |

Table 4.2: Cascade ST experiment results

# Chapter 5

# End-to-End Speech-to-text Translation

## 5.1 Motivation for End-to-End Approach

One of our main insights from the cascaded ST experiments was that the Arabic ASR transcription acts as an information bottleneck between the ASR and MT steps, and this is only exacerbated by the addition of the unsupervised MT step in the 3-step cascaded ST, causing the BLEU score to decrease from 14.7 to 14.3.

One way to get around this information bottleneck is to use a single end-to-end neural model instead of a cascaded approach. This would involve translating from Arabic speech to English text by learning the English text directly from the Arabic speech waveforms. Since the same computational model, the neural network, has achieved state-of-the-art performance for both major parts of the speech-to-text translation task (ASR and MT), it may be possible to learn speech-to-text translation end-to-end. However, a major challenge is that there is limited supervised data for training an end-to-end model, which requires a large parallel corpus of Arabic Speech to English text sentence pairs.

To the best of our knowledge, no such corpus of significant size exists where the Arabic dialect is Modern Standard Arabic (MSA). The closest contender to such a corpus is the “The CALLHOME Egyptian Arabic Speech Translation Corpus” [21], which is introduced in chapter 2. Based on the amount of data needed for Spanish-English speech-to-text translation [34], this corpus is too small for training an Arabic-English speech-to-text translation system. Perhaps more significantly, the dataset is not a good fit for our use case because the Egyptian dialect differs substantially from MSA, and is therefore unlikely to work well for MSA. With no dataset consisting of MSA speech utterances and their corresponding English text translations, it is difficult to make progress on the end-to-end approach.

We also considered constructing a training dataset. The end-to-end model may work well if we were to construct a parallel dataset comparable in size to the one used for Spanish-English speech-to-text translation (on the order of 170 thousand utterances) [34]. However, there are caveats. The first caveat is that constructing a high quality speech-to-text translation dataset is time consuming and costly. For an estimate of the costs of collecting enough parallel data for my task, we studied a paper that investigates English to Spanish speech-to-text translation [34]. The paper reported a total cost of $15,665 for creating a parallel dataset with 170 thousand utterances. We expect that this cost is a lower bound for the cost of constructing an English-Arabic speech-to-text dataset. Because English and Arabic are less similar than English and Spanish, the Arabic-English task would require more training data to achieve similar performance. Also, good crowd source workers may be more costly since Arabic speakers are less represented on crowd sourcing websites such as Amazon Mechanical Turk. The United Arab Emirates is the only Arab country among the top 20 most represented countries by number of Amazon Mechanical Turk workers [24]. Additionally, even after creating a training dataset, it is not certain that it would help. We could only test this after the expenditure of time and effort to create the dataset. Finally, even if having such a dataset enables improved translations from MSA speech to English text, the resulting translation method would not be generalizable to other language pairs.

While we did not collect a training dataset large enough for end-to-end ST, we did end up collecting a small training dataset with 19,941 parallel utterances, as described earlier in the Chapter 2. Later, in Chapter 4, we use a modified version of this training set to train a semi-supervised transcribed Arabic to MSA model in as part of a 3-step cascade ST system, but we have not explored its use for end-to-end ST due to it being an order of magnitude smaller than the amount of data used for end-to-end Spanish-English translation.

## 5.2 End-to-End ST with Synthetic Parallel Data Augmentations

In the absence of a large enough supervised dataset for training an end-to-end speech-to-text translation model, we consider approaches to train an end-to-end model without parallel data. One way to achieve this is by synthesizing parallel data that closely resembles natural parallel data, and then using this synthetic data to train the end-to-end model. We explore and compare two main approaches to synthesizing parallel data: 1) parallel data with synthetic Arabic speech, and 2) parallel data with synthetic English text. A variation of the first approach has been shown to work well for end-to-end Spanish-English speech-to-speech translation [35].

### 5.2.1 End-to-End ST with Synthetic Arabic Speech

#### Method

We use a text-to-speech (TTS) model to synthesize Arabic speech from an Arabic-English parallel text dataset to create pseudo-parallel data made up of synthetic speech and its corresponding English text. Specifically, we used the WaveNet TTS model [36] via Google Cloud Platform's Text-to-Speech API, and used all three WaveNet voices that were available for Arabic at the time. Two of these voices were male, and one female. To maximize the amount of variation in the synthesized speech, we used each of these three voices an equal number of times, and ran the TTS model on a random subset of the Arabic sentences from the 41 million pair Arabic-English parallel text dataset. In total, we synthesized 500,000 utterances of Arabic speech to form a pseudo-parallel training dataset.

We then use this pseudo-parallel data to train an end-to-end speech-to-text translation model. To train the model, we use the ESPNet speech processing toolkit [37]. The model use is an encoder-decoder transformer architecture with a speech encoder and a translation decoder as described in [38]. In the preprocessing stage, we augment the training dataset via speed-perturbation of the speech. We perturb the speed of the speech by multipliers of 0.9, 1.0 (no change), and 1.1. We perform these perturbations on the entire training dataset, resulting in a three-fold increase in its size from 500,000 to 1,500,000 utterances. We initialize the speech encoder with the parameters of the trained Arabic ASR model. Finally, since we have the MSA sentences which we used to synthesize Arabic speech, we incorporate these MSA sentences to train the model with a multi-task learning objective.

This removes the information bottleneck without deviating from the low-resource setting. In fact, this approach needs the same types of datasets as the 2-step cascaded ST approach. In particular, the WaveNet TTS model used in this approach requires an Arabic speech to Arabic text parallel dataset, as does the ASR model in the cascade ST. This approach also uses a parallel Arabic-English text dataset, which is also used to train an MT model for the cascade ST.

However, this approach and the cascade ST experiments use a different amount of training data, making it more difficult to draw a fair comparison between the two approaches. Specifically, we only used 500,000 sentence pairs from the 41 million Arabic-English sentence pairs used to train the MT model of the cascaded ST. Similarly, the amount of data used to train the Arabic Wavenet model is undisclosed, and is likely more than the 375,103 utterances our cascade ST's ASR model is trained on. We keep this limitation in mind as we discuss the results in the next section.

### Results

Training an end-to-end ST model with 500,000 synthetic parallel sentences caused the model to overfit to the synthetic speech. This can be seen in the loss curve in Figure 5-1. In an attempt to amend the overfitting, we performed several versions of this experiment, introducing various forms of regularization. First, we introduced speed-perturbations because all of the synthetic speech utterances were generated by the text-to-speech model with the same speed. Similarly, we used the SpecAugment data augmentation method [39] to further diversify the speech data. Next, we initialized the encoder parameters with the parameters of the pretrained Arabic ASR model used in the cascaded ST chapter. We also attempted to freeze encoder layers to prevent overfitting to the synthetic speech. However, these attempts did not result in significant differences in translation quality.

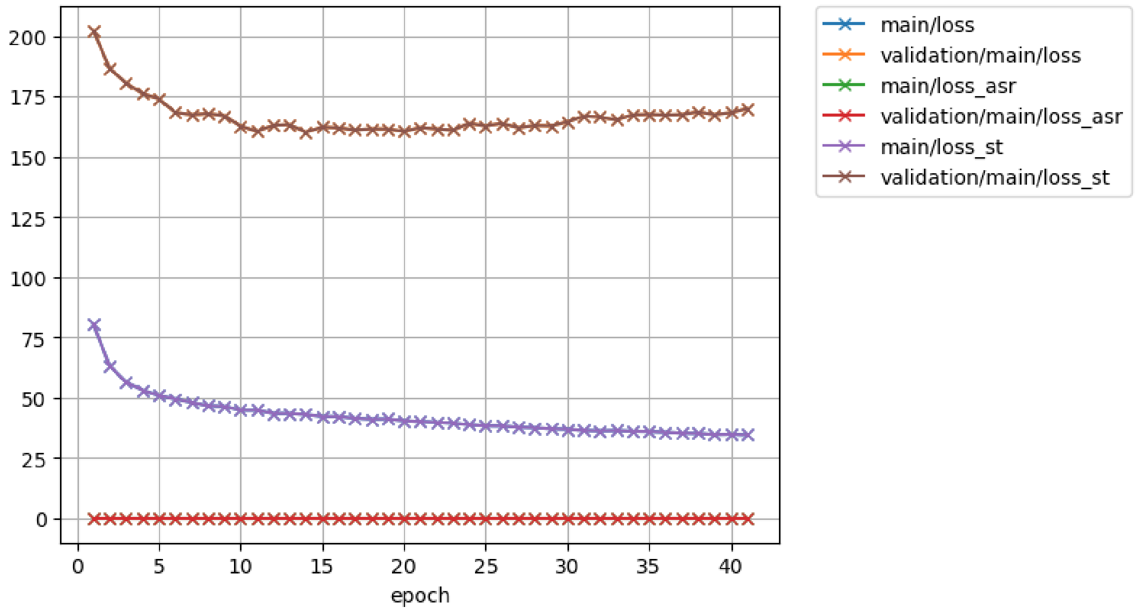


Figure 5-1: End-to-end ST with Synthetic Speech Loss Curve.

### 5.2.2 End-to-End ST with Synthetic English Text

#### Motivation

Training an end-to-end ST model with synthetic speech caused the model to overfit to the synthetic speech and therefore perform poorly on natural speech. One way to avoid training the model on synthetic speech is to construct pseudo-parallel data with natural Arabic speech and synthetic English text, instead of using synthetic Arabic speech and natural English text.

### Method

To create a pseudo-parallel dataset with natural Arabic speech and synthetic English text, we start with an ASR dataset (parallel Arabic speech and text) and translate the Arabic Text into English using a pretrained Arabic-English MT model. In particular, we use the MGB-2 Arabic ASR dataset, which consists of 375,103 Arabic utterances and transcriptions. Then, to generate the English pseudo-translations, we use the same Arabic-English MT model used in the cascaded ST approach, which is trained on 40 million Arabic-English sentence pairs.

A limitation of this approach to creating pseudo-parallel data is that the Arabic-English MT model was trained on MSA, but applied on ground truth transcripts of Arabic speech, which differ in style from MSA. This limitation is similar to that of the 2-step cascaded ST, which uses the same MT model. However, in the cascaded ST, the MT model is applied on ASR output rather than ground truth transcripts. This makes the limitation more severe in the case of the cascaded ST because, in addition to differing from MSA in style, the ASR output can also be erroneous, causing additional error propagation. As a result, since this limitation is less severe for this approach relative to the cascaded ST approach, it is actually an advantage when comparing the two.

### Experiments and Results

We trained the end-to-end ST with synthetic English text in the same way as in the previous section for end-to-end ST with synthetic speech, but replacing the 500,000 pairs of synthetic speech pseudo-parallel data with the 375,103 pairs of MGB-2 utterances with their corresponding English pseudo-translations. The resulting model achieves a BLEU score of 14.21 on the test dataset. This is the best performance of an end-to-end ST approach so far, and comes close to the 14.7 BLEU score achieved by the 2-step cascaded ST. This comparison of BLEU scores is a fair way to evaluate the two approaches because both use the exact same training datasets and amounts of data. This result suggests that with some improvements, this end-to-end ST with synthetic English text approach has the potential to outperform the cascaded ST. We explore a strategy for improving this approach next.

### 5.2.3 Filtering Pseudo-parallel dataset via Dropout-based Uncertainty-driven Self-Training

The pseudo-parallel data with synthetic English text differs from conventional parallel data in that the translations are not guaranteed to be correct. Pseudo-parallel pairs vary in quality depending on the accuracy of the translation that formed them. If we are able to quantify the quality of these translations, we can filter out the lowest quality pseudo-parallel pairs, increasing the average quality of the pseudo-parallel dataset. However, this does also decrease the size of the training dataset, creating a trade off between pseudo-translation quality and dataset size.

The standard way of quantifying the quality of a translation is by calculating a BLEU score against a ground truth translation. However, we do not have such ground truth translations, as they are precisely what we are attempting to create. We therefore need a way to quantify the quality of the translations in the absence of ground truth translations. Recent work introduces DUST (Dropout-based Uncertainty-driven Self-Training) [40], providing a method for estimating the confidence of a translation model on a given translation.

First, we translate the MGB-2 ground truth Arabic transcript to generate a pseudo-translation for each sentence. For clarity, we will refer to this translation as the original pseudo-translation. Our goal is to quantitatively evaluate how confident the MT model is in its original pseudo-translation for each sentence. Next, we modify the Arabic-English MT model by enabling the dropout layers during decoding. With dropout layers activated, we translate the MGB-2 ground truth Arabic transcripts to generate three additional pseudo-translations for each sentence. As dropout adds randomness to the previously deterministic decoding process, these three dropout pseudo-translations are not necessarily identical to each other, or to the original pseudo-translation.

We evaluate as follows. If the dropout pseudo-translations are all similar to the original pseudo-translation, i.e. if the model repeatedly produces a similar translation to the original pseudo-translation despite the activation of dropout, then we can conclude that the model had high confidence in the original pseudo-translation. To quantify this similarity, we calculate the character-based edit distance between each of the pseudo-translations and the original pseudo-translation, resulting in three measures of difference from the original pseudo-translation. Let $\mathcal{E}$ be the set of these three edit distances. Then, $max(\mathcal{E})$ is a measure of the uncertainty (or an inverse measure of confidence) of the model for the given translation. We then omit sentences from the training data if their $max(\mathcal{E})$ exceeds an uncertainty threshold. Lower uncertainty thresholds result in smaller training datasets because more data is filtered out. This has the effect of omitting training data that the Arabic-English MT model is less confident about, thus improving the average quality of sentence pairs in the dataset, but reducing its size.

### Experiment Results and Analysis

We ran three experiments with different confidence thresholds, and report the BLEU scores in Table 5.1. Setting a threshold of 0.7 filtered out a small proportion of the dataset, so the resulting BLEU score was not changed significantly relative to the model without DUST filtering. Lowering the threshold to 0.4 caused a decrease in BLEU score relative to the model without DUST filtering. We expect that this is because the decrease in performance due to the reduction of the dataset size was larger than the benefit of improved average sentence pair quality.

Additionally, the authors of the DUST paper [40] recommended a threshold of 0.4. However, this threshold causes a 43% reduction in the size of our dataset. Therefore this threshold is too severe for our use case. Our setting differs from the setting described in the DUST paper. The authors of the DUST paper used the DUST filtering method to create pseudo-parallel data to augment a statically sized supervised training dataset, so setting such a low threshold and filtering out a large proportion of the data is acceptable for their use case, as they have enough training data to train

| DUST Threshold | Train Dataset Size | % of Data Kept | BLEU |
|---|---|---|---|
| No DUST filtering (Baseline) | 375,103 | 100% | 14.7 |
| 0.7 | 326,235 | 87% | 14.6 |
| 0.4 | 213,193 | 57% | 11.9 |

Table 5.1: End-to-end ST with DUST filtering experiment results.

the model regardless of how few augmented pseudo-parallel sentences they have.

Therefore, it is possible that if we started with a larger dataset before applying the DUST filtering, the decrease in dataset size might have been less influential. This would allow us to use a threshold as low as 0.4.

## 5.3 Chapter Summary

In this chapter, we proposed two approaches for training end-to-end ST models, one with synthetic Arabic speech and one with synthetic English text. We showed that the approach with synthetic English text can achieve a BLEU score of 14.3, which is comparable to the cascaded ST baseline. Finally, we proposed a method for improving the end-to-end ST with synthetic English text by filtering the synthetic pseudo-parallel training data using the DUST method as an unsupervised measure of its quality.

# Chapter 6

# Conclusions

## 6.1 Summary of Contributions

In this thesis, we conducted research on Arabic-English speech-to-text translation, focusing on MSA. Our main contributions consist of the following:

- We construct a three-way parallel MSA-English speech-to-text translation dataset, consisting of 19,941 training sentences and 6692 test sentences.
- We propose a novel approach to augmenting cascaded ST with an unsupervised MT intermediate step to align the output of the ASR model with the input of the MT model. We showed that despite the increase in BLEU score of the intermediate-step, the 2-step cascaded ST still outperformed the 3-step approach.
- We propose two approaches for training end-to-end ST models, one with synthetic Arabic speech and one with synthetic English text. We show that the approach with synthetic English text can achieve comparable translation quality to the cascaded ST approach.
- Finally, we propose a method for improving the end-to-end ST with synthetic English text by filtering the synthetic pseudo-parallel training data using the DUST method as an unsupervised measure of its quality.

## 6.2 Future Work

We hope that our work encourages further work in Arabic-English speech translation, and more generally, improving translation models for low-resource language pairs. We see a few possible directions for future work:

- We plan to release our Arabic-English speech-to-text translation train and test datasets to the public, and encourage its use for training and evaluating translation models

- As the end-to-end ST with synthetic English text approach achieved comparable results to the cascaded ST approach, we believe that this is a promising direction. We encourage work that builds on this approach. For example, we encourage research on evaluating and improving the quality of synthetic pseudo-parallel data to improve the resulting ST model it is trained on

- While we explored a semi-supervised approach for the 3-step cascade approach, we did not explore semi-supervised end-to-end ST. This could be an effective way of building on the end-to-end approach.

- We encourage future work that extends ideas in this thesis to speech translation for dialectal Arabic

# Bibliography

[1] Glenn Flores. Language barriers to health care in the united states. *New England Journal of Medicine*, 355(3):229–231, 2006. PMID: 16855260.

[2] Imane Guellil, Houda Saâdane, Faical Azouaou, Billel Gueni, and Damien Nouvel. Arabic natural language processing: An overview. *Journal of King Saud University-Computer and Information Sciences*, 33(5):497–507, 2021.

[3] Awatef Zughoul, Muhammad ; Abu-Alshaar. English/arabic/english machine translation: A historical perspective. *Meta*, 50(3):1022–1041, 2005.

[4] Reut Tsarfaty, Djamé Seddah, Yoav Goldberg, Sandra Kuebler, Yannick Versley, Marie Candito, Jennifer Foster, Ines Rehbein, and Lamia Tounsi. Statistical parsing of morphologically rich languages (SPMRL) what, how and whither. In *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 1–12, Los Angeles, CA, USA, June 2010. Association for Computational Linguistics.

[5] Zainab Igaab and Israa Kareem. Affixation in english and arabic: A contrastive study. *English Language and Literature Studies*, 8:92, 02 2018.

[6] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473, 2015.

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

[8] Steve Renals, Nelson Morgan, Hervé Bourlard, Michael Cohen, and Horacio Franco. Connectionist probability estimators in hmm speech recognition. *IEEE transactions on speech and audio processing*, 2(1):161–174, 1994.

[9] Nelson Morgan and Herve Bourlard. Continuous speech recognition. *IEEE signal processing magazine*, 12(3):24–42, 1995.

[10] Herve A Bourlard and Nelson Morgan. *Connectionist speech recognition: a hybrid approach*, volume 247. Springer Science & Business Media, 2012.

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

[12] Salima Harrat, Karima Meftouh, and Kamel Smaili. Machine translation for arabic dialects (survey). *Information Processing & Management*, 56(2):262–273, 2019.

[13] Hassan Sajjad, Fahim Dalvi, Nadir Durrani, Ahmed Abdelali, Yonatan Belinkov, and Stephan Vogel. Challenging language-dependent segmentation for Arabic: An application to machine translation and part-of-speech tagging. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 601–607, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[14] Arfath Pasha, Mohamed Al-Badrashiny, Mona T Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan Roth. Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of arabic. In *Lrec*, volume 14, pages 1094–1101. Citeseer, 2014.

[15] Ahmed Abdelali, Kareem Darwish, Nadir Durrani, and Hamdy Mubarak. Farasa: A fast and furious segmenter for arabic. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Demonstrations*, pages 11–16, 2016.

[16] Nizar Habash, Ryan Roth, Owen Rambow, Ramy Eskander, and Nadi Tomeh. Morphological analysis and disambiguation for dialectal arabic. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 426–432, 2013.

[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, 2016.

[18] Oliver Bender, Evgeny Matusov, Stefan Hahn, Sasa Hasan, Shahram Khadivi, and Hermann Ney. The rwth arabic-to-english spoken language translation system. In *2007 IEEE Workshop on Automatic Speech Recognition Understanding (ASRU)*, pages 396–401, 2007.

[19] Cameron S. Fordyce. Overview of the iwslt 2007 evaluation campaign. In *In IWSLT*, 2007.

[20] Murat Akbacak, Horacio Franco, Michael Frandsen, Sasa Hasan, Huda Jameel, Andreas Kathol, Shahram Khadivi, Xin Lei, Arindam Mandal, Saab Mansour, Kristin Precoda, Colleen Richey, Dimitra Vergyri, Wen Wang, Mei Yang, and Jing Zheng. Recent advances in sri's iraqcomm$^{\text{TM}}$ iraqi arabic-english speech-to-speech translation system. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4809–4812, 2009.

[21] Gaurav Kumar, Yuan Cao, Ryan Cotterell, Chris Callison-Burch, Daniel Povey, and Sanjeev Khudanpur. Translations of the callhome egyptian arabic corpus for

conversational speech translation. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, Lake Tahoe, US, December 2014.

[22] Changhan Wang, Anne Wu, and Juan Pino. Covost 2 and massively multilingual speech-to-text translation, 2020.

[23] 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, July 2002. Association for Computational Linguistics.

[24] Djellel Difallah, Elena Filatova, and Panos Ipeirotis. Demographics and dynamics of mechanical turk workers. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pages 135–143, 2018.

[25] Linhao Dong, Shuang Xu, and Bo Xu. Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5884–5888. IEEE, 2018.

[26] Shigeki Karita, Nanxin Chen, Tomoki Hayashi, Takaaki Hori, Hirofumi Inaguma, Ziyan Jiang, Masao Someki, Nelson Enrique Yalta Soplin, Ryuichi Yamamoto, Xiaofei Wang, et al. A comparative study on transformer vs rnn in speech applications. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 449–456. IEEE, 2019.

[27] Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R Hershey, and Tomoki Hayashi. Hybrid ctc/attention architecture for end-to-end speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1240–1253, 2017.

[28] Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. Unsupervised neural machine translation. In *International Conference on Learning Representations*, 2018.

[29] Guillaume Lample, Ludovic Denoyer, and Marc'Aurelio Ranzato. Unsupervised machine translation using monolingual corpora only. *CoRR*, abs/1711.00043, 2017.

[30] Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049, 2018.

[31] Alexis Conneau and Guillaume Lample. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems*, 32:7059–7069, 2019.

[32] Nizar Habash and Fatiha Sadat. Arabic preprocessing schemes for statistical machine translation. In *HLT-NAACL*, 2006.

[33] Amjad Almahairi, Kyunghyun Cho, Nizar Habash, and Aaron C. Courville. First result on arabic neural machine translation. *CoRR*, abs/1606.02680, 2016.

[34] Matt Post, Gaurav Kumar, Adam Lopez, Damianos Karakos, Chris Callison-Burch, and Sanjeev Khudanpur. Improved speech-to-text translation with the fisher and callhome spanish–english speech translation corpus. In *Proc. IWSLT*, 2013.

[35] Ye Jia, Ron J Weiss, Fadi Biadsy, Wolfgang Macherey, Melvin Johnson, Zhifeng Chen, and Yonghui Wu. Direct speech-to-speech translation with a sequence-to-sequence model. *Proc. Interspeech 2019*, pages 1123–1127, 2019.

[36] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio, 2016.

[37] Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai. ESPnet: End-to-end speech processing toolkit. In *Proceedings of Interspeech*, pages 2207–2211, 2018.

[38] Hirofumi Inaguma, Shun Kiyono, Kevin Duh, Shigeki Karita, Nelson Yalta, Tomoki Hayashi, and Shinji Watanabe. ESPnet-ST: All-in-one speech translation toolkit. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 302–311, Online, July 2020. Association for Computational Linguistics.

[39] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. Specaugment: A simple data augmentation method for automatic speech recognition. *Interspeech 2019*, Sep 2019.

[40] Sameer Khurana, Niko Moritz, Takaaki Hori, and Jonathan Le Roux. Unsupervised domain adaptation for speech recognition via uncertainty driven self-training. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6553–6557. IEEE, 2021.