End-to-End Speech Translation of Arabic to English Broadcast News

Fethi Bougares
Le Mans Université - France
`fethi.bougares@univ-lemans.fr`

Salim Jouili
Elyadata - Tunisia
`salim.jouili@elyadata.com`

Abstract

Speech translation (ST) is the task of directly translating acoustic speech signals in a source language into text in a foreign language. ST task has been addressed, for a long time, using a pipeline approach with two modules : first an Automatic Speech Recognition (ASR) in the source language followed by a text-to-text Machine translation (MT). In the past few years, we have seen a paradigm shift towards the end-to-end approaches using sequence-to-sequence deep neural network models. This paper presents our efforts towards the development of the first Broadcast News end-to-end Arabic to English speech translation system. Starting from independent ASR and MT LDC releases, we were able to identify about 92 hours of Arabic audio recordings for which the manual transcription was also translated into English at the segment level. These data was used to train and compare pipeline and end-to-end speech translation systems under multiple scenarios including transfer learning and data augmentation techniques.

1 Introduction

End-to-end approach to speech translation is gradually replacing the cascaded approach which consists of transcribing the speech inputs with an ASR system, and translating the obtained transcription using a text-to-text MT system. For instance, and for the first time, the winning system in the IWSLT 2020 TED English-to-German speech translation shared task was an end-to-end system (Ansari et al., 2020). Despite this positive result, the end-to-end approach is used on a limited scale due to the lack of labeled data. Indeed, data scarcity is today the major blocker for the widespread adoption of the end-to-end models. Taking this into consideration, recent works have focused on developing speech translation corpora. Joint efforts in this direction have allowed us to collect a significant quantity and good quality of speech translation corpora. Not surprisingly, speech translation corpus development has started for well-resourced languages including English and some European languages. In (Kocabiyikoglu et al., 2018), the 236 hours English→French ST Augmented LibriSpeech were released. Shortly after, (Di Gangi et al., 2019) released the MUST-C corpus including few hundreds (385 to 504 hours) of parallel ST data of TED talks translations from English to eight European languages. At the same time, the Europarl-ST (Iranzo-Sánchez et al., 2019) was released with translations between 6 European languages, with a total of 30 translation directions. While all the previous resource development effort has focused on well-resourced languages, the most recent published corpora CoVoST (Wang et al., 2020a) and CoVoST2 (Wang et al., 2020b). These latter works released a large-scale Multilingual Speech-To-Text Translation Corpus covering translations from 21 languages to English and from English to 15 languages. Although the Arabic-English is one of the language pairs covered by the CoVoST2 corpus, the authors consider it as a low-resource pair. In fact, CoVoST2 corpus contained only 6 hours of prepared speech uniformly splitted between train, dev and test sets. In this paper, we conduct a series of experiments to present the first results of Arabic to English End-to-End Broadcast News Speech translation.

This paper is organized as follows: section 2 presents Arabic-to-English speech translation related works. Section 3 gives details about the source of our training data and the method we have used to extract theses data. In section 4, we present our experimental setup as well as the used toolkits to train ou models. Sections 5 and 6 provides details about the pipeline and end-to-end speech translation systems, respectively. Section 7, gives a brief discussion and results analysis. Finally, section 8 concludes the findings of this study and discusses future work.

2 Related works

Arabic-English (AR-EN) is one of the most studied language pair in the context of Speech Recognition and Machine Translation. For instance, this language pair was integrated in several evaluation campaigns and projects including the International Workshop on Spoken Language Translation (IWSLT) and DARPA's Babylon project. These earlier projects have built on the traditional pipelined architecture integrating speech recognition system in the source language followed by machine translation from the transcript to the target language. In IWSLT, the speech translation task was introduced for the first time in 2006. The IWSLT06 (Paul, 2006) translation campaign was carried using either the manual or the automatic transcription of speech input in the travel domain. This translation task was renewed for several years using always the pipeline approach.

Pipeline architecture was also used by BBN in the context Babylon project (Stallard et al., 2011). They developed the TransTalk system including a pipeline of ASR and MT systems in both directions (AR↔EN). More recently, but still with the same approach, QCRI presented their live Arabic-to-English speech translation system in (Dalvi et al., 2017). The system is a pipeline of a Kaldi-based Speech Recognition followed by a Phrase-based/Neural MT system. Recently, there has been a shift to the most recent end-to-end approach without the intermediate step of transcribing the source language. Indeed, IWSLT 2018 was the first time where organizers drooped the ASR task and participants needed to develop an end-to-end speech translation systems. End-to-end speech translation has shown its effectiveness for multiple languages and in multiple scenarios. It becomes now a well-established task in IWSLT evaluation campaign were multiple shared taks are proposed to assess Spoken Language Translation (SLT) systems for many language pair in several settings. Despite the great interest being shown to the end-to-end approach for speech translation task, we were able to identify only one recent work by (Wang et al., 2020b) including Arabic-English language pair limited to a corpus of 6 hours. We are also aware of the IWSLT 2022 Dialectal Speech Translation[1] task which, unlike

this work, focuses on Tunisian-to-English speech translation (Zanon Boito et al., 2022; Yan et al., 2022; Yang et al., 2022).

3 Training Data

Whatever the chosen architecture for Speech translation system (pipeline or end-to-end), it requires a large amount of manually annotated training data that might be hard to obtain for multiple language pairs. For the Arabic-English language pair, a large amount of training data for ASR and MT was manually annotated in the framework of the DARPA's Global Autonomous Language Exploitation (GALE) project (Cohen, 2007). This huge amount of work was done for the purpose of making foreign languages speech and text accessible to English-only speakers through the development of automatic speech recognition and machine translation systems.

In this respect, Arabic broadcast news and conversation speech were collected from multiple sources, then annotated under the supervision of Linguistic Data Consortium (LDC). Audio corpora and their transcripts are separately released in three phases : GALE Phase 2, 3 and 4. In addition to the speech audio corpora and transcripts released to train Arabic ASR systems, LDC also made available multiple Arabic to English parallel corpora.The latter are intended to be used for training and evaluating Arabic to English MT systems. They have been developed by manually translating from a number of different sources including Arabic news-wire, discussion groups and broadcast news and conversation.

Upon closer inspection of these parallel corpora, we have found that part of the broadcast news and conversation parallel corpora were built by translating the manual transcripts released for the ASR task. Following the discovery we dug deep in the GALE speech recognition and machine translation LDC related releases and, as illustrated in Figure 1, we parsed all GALE speech recognition and machine translation corpora in order to extract a 3-way parallel corpus consisting of Arabic audio along with their Arabic transcriptions and English translation. As shown in Figure 1, for each transcribed audio file part of the GALE corpus, we extract only segments for which we were able to find an exact match between the

manual transcription, from ASR training data, and the source side of parallel corpora. Table 1 shows the amount of speech audio for which we were able to find the corresponding translation in the LDC MT related releases. We report, for each phase: 2, 3 and 4, the original size of the speech corpus in hours and the extracted subset for which the English translation had been found.

GALE Phase	Hours	#Segments
Phase 2	436h 11 min	190.510
Phase 2 ST.	59h 12 min	24.519
Phase 3	419h 03 min	195.143
Phase 3 ST.	28h 50 min	13.261
Phase 4	96h 18 min	54.787
Phase 4 ST.	4.0h 08 min	1.855
Phase 2/3/4	951h 32 min	440.440
Extracted ST.	92h 10 min	39.635

Table 1: Statistics of the original GALE Arabic to English Speech Transcription corpus and the extracted subsets for which translations are available.

All the extracted segments were afterwards aligned using timestamps information from ASR transcript and translation from MT target side. As table 1 shows, an overall Arabic to English speech translation corpus of around 92 hours was extracted. This corpus was then cleaned out by removing all the back-channel and incomplete speech segments. The final corpus is then splitted into training, development and test sets. Development and test contain segments from randomly selected broadcast audio programs. Their size is respectively 1188 and 987 segments. Development set contain broadcast News and Conversation recordings from Abu Dhabi TV, Al Alam News Channel, based in Iran and Al Arabiya. Test set is made up of broadcast News and Conversation recordings from Abu Dhabi TV, Aljazeera, Al Arabiya and Syria TV. The remaining material was used as training data for ASR, MT and ST systems.

Table 2 gives a detailed statistics of the extracted Arabic to English Speech Translation corpus, including speech duration as well as token counts for both transcripts and translations.

	Train	dev.	test
Hours	83h54	3h05	2h38
Sentences	32.099	1188	987
#AR words	606.465	22.537	18.598
#EN words	945.801	35.180	27.880

Table 2: Statistics and splits of the extracted Arabic to English Speech Translation corpus extracted from LDC ASR and MT independent releases.

4 Experimental Setup

All our experiments are built using open source toolkits with the following settings: ASR models were built using the End-to-End Speech Processing Toolkit ESPnet (Watanabe et al., 2018b). We trained an attention-based encoder-decoder architecture with an encoder of 4 VGGBLSTM layers including 1024 cells in each layer. The second and third bottom BLSTM layers of the encoder reduced the utterance length by a factor of two. We used a decoder of one 1024-dimensional BLSTM layer. For both ASR and ST speech utterance, we extracted 40 Melfilterbank energy features with a step size of 10ms and a window size of 25ms. The extracted features, we applied mean and variance normalization. MT models were built using the FAIRSEQ package (Ott et al., 2019a). We trained end-to-end word and *bpe-based* translation systems using the *"lstm_luong_wmt_en_de"* model template. This template is a standard LSTM Encoder-Decoder architecture composed of 4 stacked BLSTM layers, each with 1000 cells, and 1000-dimensional embeddings (Luong et al., 2015). Translation tasks (AST and MT) evaluation was carried out using case-sensitive BLEU metric (Papineni et al., 2002). Scores are calculated using one human reference with Moses' mteval-v14.pl script [2] applied to de-tokenized and punctuated translation output. As for ASR, systems were evaluated using Word Error Rate (WER).

5 Pipeline Speech Translation

In this section, we evaluate the pipeline approach for speech translation in two different scenarios, plausible for many language pairs, depending on the amount and the type of training data used for the development of the Speech Translation task.

1. **Constrained Scenario** : Under this scenario

[2] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic

Audio

Source Language (AR)

من المتوقع أن يلقي الرئيس الأمريكي جورج بوش خطابا مساء اليوم
يستعرض فيه استراتيجية الولايات المتحدة الجديدة بشأن العراق
غير أن الديمقراطيين الذين سيطروا لتوهم على الكونغرس هددوا بالاعتراض على هذه الخطة
وقد بدأ السيناتور الأمريكي كينيدي لإثناء الرئيس الأمريكي من إرسال هذه القوات
وقال كينيدي إنه سيطرح تشريعا
نظام الحكم ملكي دستوري والعاصمة عمان
متوسط الدخل السنوي للفرد ألفان ومائة وأربعون دولارا
الصادرات الرئيسية الفوسفاط الأسمدة والمنتجات الزراعية
موارد إضافية تحويلات المغتربين والسياحة
لمدة سبعة وأربعين عاما حكم المملكة الفتية الملك حسين بن طلال
لكن الاقتصاد لايزال في جزء منه على المساعدات الأجنبية
بهذا تنتهي نشرة الأخبار
في الختام شكرا لمتابعة إلى اللقاء

Transcription

ABUDHABI_ABUDHNEW 1389.84 1392.54 spk1 male أهلا بكم إلى نشرة الأخبار
ABUDHABI_ABUDHNEW 1392.54 1402.84 spk1 male إغتالت قوات خاصة إسرائيلية من وحدات المستجوبين ثلاثة مقاومين في مخيم جنين
ABUDHABI_ABUDHNEW 1402.84 1408.56 spk1 male كما واصل الجيش الإسرائيلي توغله في مدينة نابلس بشمال الضفة الغربية
ABUDHABI_ABUDHNEW 1408.56 1412.59 spk1 male نظام الحكم ملكي دستوري والعاصمة عمان
ABUDHABI_ABUDHNEW 1412.59 1417.44 spk1 male متوسط الدخل السنوي للفرد ألفان ومائة وأربعون دولارا
ABUDHABI_ABUDHNEW 1417.44 1422.13 spk1 male الصادرات الرئيسية الفوسفاط الأسمدة والمنتجات الزراعية
ABUDHABI_ABUDHNEW 1422.13 1426.33 spk1 male موارد إضافية تحويلات المغتربين والسياحة
ABUDHABI_ABUDHNEW 1426.33 1432.16 spk1 male لمدة سبعة وأربعين عاما حكم المملكة الفتية الملك حسين بن طلال
ABUDHABI_ABUDHNEW 1440.41 1445.49 spk1 male ضافة إلى معلجة المطالب المتنامية من أجل الإصلاح السياسي والاقتصادي والاجتماعي
ABUDHABI_ABUDHNEW 1464.54 1471.87 spk1 male رمزي المصري العربية
ABUDHABI_ABUDHNEW 1471.87 1474.09 spk1 male

Exact matching

Translation

American President George Bush is expected to give a speech this evening
in which he outlines the new American strategy on Iraq.
However, the Democrats, who have just taken control of the Congress, threatened to oppose this plan.
The American Senator Kennedy attempted to dissuade the American president from sending these troops.
Kennedy said that he will propose legislation
The number of Jordan's population reaches approximately seven million.
The government is a constitutional monarchy and the capital is Amman.
The average yearly income per person is 2,140 dollars.
The main exports are phosphate fertilizers and agricultural products.
Additional revenues are the expatriates' remittances and tourism.
For 47 years, the young kingdom was ruled by King Hussein bin Talal.
However, the economy still depends partly on foreign aid.
This is the end of the news bulletin.
In closing, thank you for watching. Goodbye.

ABUDHABI_ABUDHNEW 1408.56 1412.59 spk1 male The government is a constitutional monarchy and the capital is Amman.
ABUDHABI_ABUDHNEW 1412.59 1417.44 spk1 male The average yearly income per person is 2,140 dollars.
ABUDHABI_ABUDHNEW 1417.44 1422.13 spk1 male The main exports are phosphate fertilizers and agricultural products.
ABUDHABI_ABUDHNEW 1422.13 1426.33 spk1 male Additional revenues are the expatriates' remittances and tourism.
ABUDHABI_ABUDHNEW 1426.33 1432.16 spk1 male For 47 years, the young kingdom was ruled by King Hussein bin Talal.
ABUDHABI_ABUDHNEW 1440.41 1445.49 spk1 male However, the economy still depends partly on foreign aid.
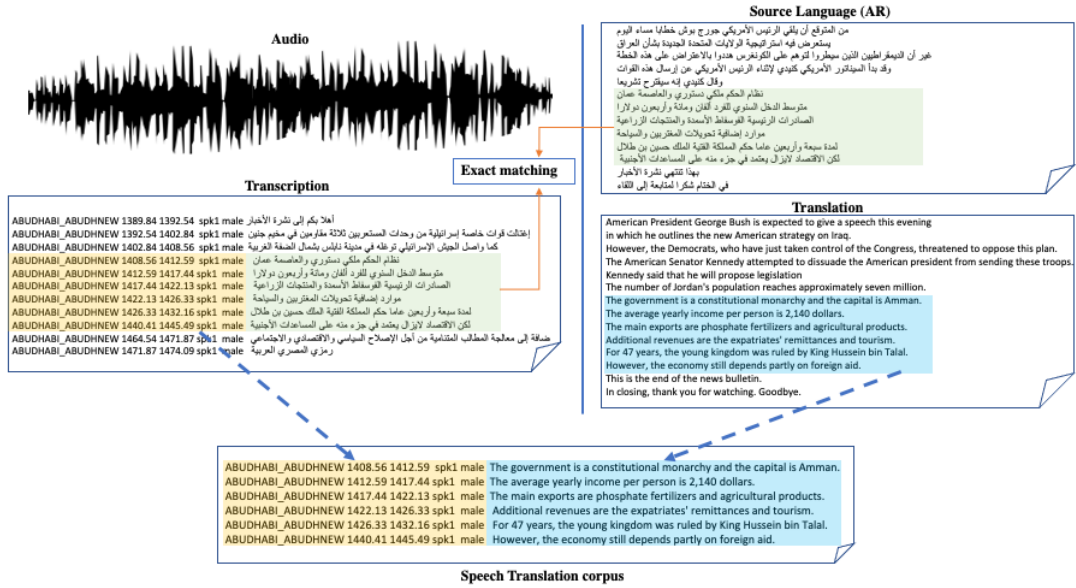
Speech Translation corpus

Figure 1: Extraction Arabic to English speech translation corpus from LDC ASR and MT independent releases.

we have access to a 3-way limited training data. This data includes speech audio files in source language their transcriptions in the source language and translations to the target language.

2. **Unconstrained Scenario** : In addition to resources from the constrained scenario, we have access to a large ASR and MT-specific resources.

As to the first scenario of the pipeline approach we only used the 3-way parallel data reported in table 2. In this instance, an end-to-end ASR module was trained using ESPnet (Watanabe et al., 2018a) toolkit on the speech audio files from Table 2 and their corresponding transcripts. In the Unconstrained Scenario, however, ASR module was trained using the totality of the GALE Phase 2, 3 and 4 ASR data reported in Table 1.

ASR System	dev	test
ASR_*Const*	20.90	21.90
ASR_*UnConst*	13.10	14.60

Table 3: ASR WER (in %) on the dev and test sets.

Table 3 presents the results of ASR system under both constrained and unconstrained scenarios. As shown in the Table 3, using a training set of around 84h of manually transcribed broadcast news and conversation, we obtained a WER of 20.90% and 21.90% on dev and test sets,

respectively. Not surprisingly, WER has been significantly improved with the use of the complete GALE training data[3] (row ASR_*UnConst*) to achieve **13.10%** and **14.60%** on dev and test sets, respectively.

As previously stated, within the pipeline ST framework, the output of the ASR module is automatically translated to the target language using the MT module. The MT module is also an end-to-end system trained using Fairseq toolkit (Ott et al., 2019b) under both constrained (MT_*Const*) and unconstrained (MT_*UnCons*) scenarios. Table 4 reports the BLEU scores of the translation output by varying ASR module condition while fixing MT module constrained to speech translation data composed of the transcripts along with their corresponding English translation from Table 2.

Pipeline ST System	dev	test
MT_*Const_ASR_*Const	19.03	15.96
MT_*Const_ASR_UnConst*	20.69	16.58
MT_*Const_ref_Transc*	22.31	18.30

Table 4: Case-sensitive tokenized and single-reference BLEU scores (in %) of the pipeline speech translation system with the constrained MT module.

The first row in table 4 (MT_*Const_ASR_*Const) gives the BLEU

---

[3]We have taken particular care to remove dev and test data before using GALE corpora to train the ASR system.

score when the MT constrain module translates the output of a constrained ASR system (row ASR_*Const* from Table 3). In this case, a BLEU score of 19.03% and 15.96% is respectively achieved on dev and test sets. The second row in the same table (MT_*Const_ASR_UnConst*) shows the BLEU score when the ASR module is under the unconstrained condition, i.e. output from the system ASR_*UnConst* in Table 3 are used as input to the MT system. As expected, when it comes to translating a higher transcription quality, the translation quality is better and the BLEU score is increased by 1.66 and 0.62 BLEU points on dev and test sets, respectively. The last row of table 4 (MT_*Const_ref_Transc*) simulates the situation where we have access to a perfect transcripts in the source language. In this case, translation quality is further improved reaching 22.31 BLEU points on dev set and 18.30 points on test set.

In a similar vein, table 5 presents results in settings where MT module is no longer constrained to speech translation data. Indeed, additional Arabic to English Bilingual text from GALE LDC releases are used to train the unconstrained MT module [4]. This unconstrained MT module, was used to run several experiments using various input conditions similar to what we did within the constrained condition. The results of these experiments are presented in Table 5. The first row (MT_*UnConst_ASR_Const*) sets out the BLEU score when the unconstrained MT module translates the output of the constrained ASR (first row in table 3). Compared to using the constrained MT system, a considerable improvement of 12.84 (from 19.03 to 31.87) and 8.26 (from 15.96 to 24.22) BLEU points is achieved on dev and test sets, respectively.

As we have seen above, translation quality is further improved when the input to the translation module is of a higher quality generated by the unconstrained ASR system (row MT_*UnConst_ASR_UnConst*). This allows to reach a dev and test BLEU scores of 36.48 and 25.80 respectively. As expected, the BLEU score is even better when it comes to translate the reference transcription (MT_*UnConst_ref_Transc*) as shown in the last row of Table 5. In the latter case,

we achieved a dev set BLEU score of 39.51 and a test set BLEU score of 30.60.

Pipeline ST System	dev	test
MT_*UnConst_ASR_Const*	31.87	24.22
MT_*UnConst_ASR_UnConst*	36.48	27.51
MT_*UnConst_ref_Transc*	39.51	30.60

Table 5: Case-sensitive tokenized and single-reference BLEU scores (in %) of the pipeline speech translation system with Unconstrained MT module.

6 End-to-End Speech Translation

In this section, we present and evaluate the end-to-end approach for Arabic to English speech translation task. The End-to-End system is built using the ESPnet toolkit (Watanabe et al., 2018b). We used an attention-based encoder-decoder architecture. The encoder has two VGG-like CNN blocks followed by five stacked 1024-dimensional BLSTM layers. The decoder is composed of two 1024-dimensional LSTM layers. Each VGG block contains two 2D-convolution layers followed by a 2D-maxpooling layer whose aim is to reduce both time and frequency dimension of the input speech features by a factor of 2. All our experiments are conducted using characters as target tokens.

Table 6 shows the performance of the end-to-end ST model with different training configurations.

End2End ST system	dev	test
Baseline (1)	2.58	2.23
(1) + Enc. init	12.44	9.57
(1) + Unsup ph234	23.23	18.97
(1) + Enc. Init + Unsup ph234	**24.95**	**19.09**

Table 6: Case-sensitive tokenized and single-reference BLEU score (in %) of the End-to-end AR→EN Speech Translation system with Encoder initialization and data augmentation

The first row from Table 6 shows the baseline results obtained when the end-to-end model is trained under the constrained scenario, that is when the training data is restricted to the 83h54 minutes from table 2. We can clearly see that the end-to-end model is not strong enough to compete with the cascaded model trained using the same amount of data. Indeed, the BLEU score of the

---

[4]Unconstrained MT system was trained using all GALE Arabic-English Parallel Text from 2007 to 2016.

end-to-end system on the dev set is 2.58, compared to the 19.03 points of the pipeline model. The same goes for test set where end-to-end system BLEU score is 2.23 compared to 15.96 which is obtained with cascade translation approach.

From this initial baseline and with the aim of improving the end-to-end system translation quality, we employed the well established transfer learning technique (Bansal et al., 2018) commonly referred as encoder pre-training. Indeed, using the ASR encoder of the Unconstrained ASR system (row ASR_*UnConst* in 3) to initialize the parameters of the ST encoder greatly improves the performance of end-to-end ST networks. The results of the encoder pre-training are shown in the second row ((1) + Enc. init) of table 6. As a result, we observed a strong effect reflected by the substantial improvement in the BLEU score: +9.86 and +7.34 BLEU score for dev and test sets, respectively.

Just like the transfer learning via encoder pre-training approach, data augmentation is proven to enhance end-to-end speech translation quality. It is carried out using synthetic data which is generated by automatically translating the transcripts of an ASR corpora in the source language. Herein, we used the unconstrained NMT system (MT_*UnConst*) of table 5 in order to translate the Arabic GALE transcripts provided in table 1. Incomplete and back-channel speech segments were filtered out from the generated translations. All in all, we were able to create the synthetic corpus of 795 hours of Arabic to English speech translation corpus detailed in table 7.

	Hours	#Sent.	#AR	#EN
Gale Synth.	795	314.167	6.1M	9.1M

Table 7: Statistics of the synthetic Ar-En ST corpus.

These synthetic data are thereafter used as additional data to train the end-to-end ST system. The results of this data augmentation experiment are highlighted in Table 6 (row (1) + unsup ph234). As we can see from the obtained results synthetic training data boosts up the end-to-end ST system to achieve a BLEU scores of **23.23** points and **18.97** points for dev and test sets, respectively.

Both encoder pre-training and data augmentation are shown to be helpful improving signifi-

cantly the ST baseline. We also experimented using both methods at the same time. The last row of the same table presents the results of the end-to-end speech translation trained with data augmentation using synthetic data from Table 7, and encoder hyperparameters initialization from the ASR_*UnConst* system presented in Table 3. By applying these two methods together, we were able to reach a BLEU scores of **24.95** and **19.09** points for dev and test sets, respectively. These end-to-end speech translation results are to be compared to pipeline results shown in row MT_*UnConst_ASR_UnConst* of table 5.

7 Discussion and analysis

Despite the improvements brought by transfer learning and data augmentation technics, the best results are still obtained using cascade architecture. We believe that this performance gap can be partly explained by the fact that end-to-end system was trained using only a small amount ($\sim$ 84 hours) of real speech translation corpus.

Based on the results of previous works from (Liu et al., 2019), the end-to-end ST models are known as an effective means of circumventing the error-propagation problem faced by the conventional pipeline system. Indeed, every involved component in the traditional pipeline approach produces errors, which are propagated through the cascade and lead to compounding follow-up errors. In order to assess the ability of our end-to-end system to overcome this error-propagation pattern, we selected some translation examples where pipeline system fails due to this problem and we checked the translation output of the end-to-end system. Example from table 8 shows a translation error caused by the propagation of transcription errors occurred at the end of the segment (text in bold ASR output row). The end-to-end system, however, relies on the source speech signal and translates correctly the same part of the input.

In addition to this error-propagation problem we have found that end-to-end system is sometimes penalized although its translation is correct. Table 9 presents and example where both systems output correct translation but BLEU score is better with pipeline system. The thing might happen for pipeline system as well, but we believe that end-

ASR output	أنا قلت أنا قلت بشكل واضح إنه هناك بعض الإتجهات المتصاعدة في العراق
ASR reference	أنا قلت أنا قلت بشكل واضح إنه هناك بعض الإتجهات المتطرفة داخل فتح
Pipeline MT	I said, I said clearly that there are some escalating trends in Iraq.
E2E MT	And now I said I said that there are some interpretations of the institutions of the Fatah movement.
MT reference	I said, I clearly said that there are, there are some extremist currents within the Fatah movement.

Table 8: Example of error-propagation problem with pipeline speech translation system.

ASR output	إسرائيل هي التي ترفض الأعتراف بحماس أليس كذلك
ASR reference	إسرائيل هي التي ترفض الأعتراف بحماس أليس كذلك
Pipeline MT	Israel is the one that refuses to recognize Hamas, isn't it?
E2E MT	Israel refuses to recognize Hamas, isn't it?
MT reference	Israel is the one that refuses to recognize Hamas, right?

Table 9: Comparison of End-to-End and pipeline translation outputs.

to-end systems are more affected as the translation references are obtained by translated from a textual input, not from speech audio in the source language. This trend must be probed further in order to quantify its impact on the end-to-end ST system performance. We leave such investigations as future work.

8 Conclusions

In this paper, we presented the first results on Arabic-to-English end-to-end Automatic Speech Translation system. Arabic-English language pairs is one of the well-studied language pair in Natural Language Processing. Therefore, large quantities of data are made available in a wide variety of domains including ASR and MT. Starting from independent LDC releases for MT and ASR systems, we were able to extract around 92 hours of speech translation corpus composed of Arabic audio and their source transcriptions and English translation. We used this corpus to conduct speech translation experiments using a pipeline and an end-to-end Speech Translation architecture. Both methods were tested under a constrained and an unconstrained conditions. We showed that the performance gap, which is too big between the two considered approaches under the constrained condition, can be narrowed under unconstrained

condition through the use of transfer learning and data augmentation techniques. In spite of considerable improvement obtained by applying these techniques, the gap remains important and we plan to reduce it in several ways including decoder pre-training, spectrogram augmentation and Self-Supervised Learning.

9 Acknowledge

References

Ebrahim Ansari, Amittai Axelrod, Nguyen Bach, Ondřej Bojar, Roldano Cattoni, Fahim Dalvi, Nadir Durrani, Marcello Federico, Christian Federmann, Jiatao Gu, Fei Huang, Kevin Knight, Xutai Ma, Ajay Nagesh, Matteo Negri, Jan Niehues, Juan Pino, Elizabeth Salesky, Xing Shi, Sebastian Stüker, Marco Turchi, Alexander Waibel, and Changhan Wang. 2020. FINDINGS OF THE IWSLT 2020 EVALUATION CAMPAIGN. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 1–34, Online. Association for Computational Linguistics.

Sameer Bansal, Herman Kamper, Karen Livescu, Adam Lopez, and Sharon Goldwater. 2018. Pre-training on high-resource speech recognition improves low-resource speech-to-text translation. *CoRR*, abs/1809.01431.

J. Cohen. 2007. The gale project: A description and an update. In *2007 IEEE Workshop on Automatic Speech Recognition Understanding (ASRU)*, pages 237–237.

Fahim Dalvi, Yifan Zhang, Sameer Khurana, Nadir Durrani, Hassan Sajjad, Ahmed Abdelali, Hamdy Mubarak, Ahmed Ali, and Stephan Vogel. 2017. QCRI live speech translation system. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 61–64, Valencia, Spain. Association for Computational Linguistics.

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. MuST-C: a Multilingual Speech Translation Corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2012–2017, Minneapolis, Minnesota. ACL.

Javier Iranzo-Sánchez, Joan Albert Silvestre-Cerdà, Javier Jorge, Nahuel Roselló, Adrià Giménez, Albert Sanchís, Jorge Civera, and Alfons Juan. 2019. Europarl-st: A multilingual corpus for speech translation of parliamentary debates. In *CoRR*, volume abs/1911.03167.

Ali Can Kocabiyikoglu, Laurent Besacier, and Olivier Kraif. 2018. Augmenting librispeech with french translations: A multimodal corpus for direct speech translation evaluation. In *CoRR*, volume abs/1802.03142.

Yuchen Liu, Hao Xiong, Zhongjun He, Jiajun Zhang, Hua Wu, Haifeng Wang, and Chengqing Zong. 2019. End-to-end speech translation with knowledge distillation. In *CoRR*, volume abs/1904.08075.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. ACL.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019a. fairseq: A fast, extensible toolkit for sequence modeling.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019b. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*

(Demonstrations), pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Michael Paul. 2006. Overview of the IWSLT 2006 evaluation campaign. In *2006 International Workshop on Spoken Language Translation, IWSLT 2006, Keihanna Science City, Kyoto, Japan, November 27-28, 2006*, pages 1–15. ISCA.

David Stallard, Rohit Prasad, Prem Natarajan, Fred Choi, Shirin Saleem, Ralf Meermeier, Kriste Krstovski, Shankar Ananthakrishnan, and Jacob Devlin. 2011. The bbn transtalk speech-to-speech translation system. In Ivo Ipsic, editor, *Speech and Language Technologies*, chapter 3. IntechOpen, Rijeka.

Changhan Wang, Juan Pino, Anne Wu, and Jiatao Gu. 2020a. Covost: A diverse multilingual speech-to-text translation corpus. In *CoRR*.

Changhan Wang, Anne Wu, and Juan Pino. 2020b. Covost 2 and massively multilingual speech-to-text translation. In *CoRR*.

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. 2018a. ESPnet: End-to-end speech processing toolkit. In *Proceedings of Interspeech*, pages 2207–2211.

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. 2018b. Espnet: End-to-end speech processing toolkit. *CoRR*, abs/1804.00015.

Brian Yan, Patrick Fernandes, Siddharth Dalmia, Jiatong Shi, Yifan Peng, Dan Berrebbi, Xinyi Wang, Graham Neubig, and Shinji Watanabe. 2022. CMU's IWSLT 2022 dialect speech translation system. In *Proceedings of the 19th International Conference on Spoken Language Translation*, pages 298–307, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Jinyi Yang, Amir Hussein, Matthew Wiesner, and Sanjeev Khudanpur. 2022. JHU IWSLT 2022 dialect speech translation system description. In *Proceedings of the 19th International Conference on Spoken Language Translation*, pages 319–326, Dublin, Ireland (in-person and online). ACL.

Marcely Zanon Boito, John Ortega, Hugo Riguidel, Antoine Laurent, Loïc Barrault, Fethi Bougares, Firas Chaabani, Ha Nguyen, Florentin Barbier, Souhir Gahbiche, and Yannick Estève. 2022. ON-TRAC consortium systems for the IWSLT 2022 dialect and low-resource speech translation tasks. In *Proceedings of the 19th International Conference on Spoken Language Translation*, pages 308–318, Dublin, Ireland (in-person and online). Association for Computational Linguistics.