TRANSLATIONS OF THE CALLHOME EGYPTIAN ARABIC CORPUS FOR CONVERSATIONAL SPEECH TRANSLATION

Gaurav Kumar[1], Yuan Cao[1], Ryan Cotterell[1], Chris Callison-Burch[2], Daniel Povey[1], Sanjeev Khudanpur[1]

[1]Center for Language and Speech Processing & HLTCOE, Johns Hopkins University, Baltimore, USA
[2]Computer and Information Science Department, University of Pennsylvania, Philadelphia, USA

ABSTRACT

Translation of the output of automatic speech recognition (ASR) systems, also known as speech translation, has received a lot of research interest recently. This is especially true for programs such as DARPA BOLT which focus on improving spontaneous human-human conversation across languages. However, this research is hindered by the dearth of datasets developed for this explicit purpose. For Egyptian Arabic-English, in particular, no parallel speech-transcription-translation dataset exists in the same domain. In order to support research in speech translation, we introduce the Callhome Egyptian Arabic-English Speech Translation Corpus. This supplements the existing LDC corpus with four reference translations for each utterance in the transcripts. The result is a three-way parallel dataset of Egyptian Arabic Speech, transcriptions and English translations.

Index Terms— Spoken Language Translation, Speech Recognition, Machine Translation, Language Resources, Corpus Creation

1. INTRODUCTION

Translation of the output of automatic speech recognition (ASR) systems, also known as speech translation, has been the subject of research for several years now. Major programs that focused on this were VERBMOBIL, NESPOLE!, DARPA TRANSTAC, DARPA GALE and the Quaero project. The early projects were limited domain and limited vocabulary systems built to cater to machine directed or well enunciated speech. However, DARPA GALE and Quaero required large vocabulary continuous speech recognition systems with generic language models for ASR, and wide coverage SMT systems for translation. As both ASR and statistical machine translation systems have become more effective over the years, speech translation has once again become a major topic of research. The focus of the most recent project, DARPA BOLT (similar to its predecessor DARPA GALE), is to build spoken language translation (SLT) systems for spontaneous, conversational, human-human speech. In contrast to machine directed or scripted conversations (broadcast news), most conversational speech has by nature, variability in recording environment and vocal registers and a high number of disfluencies and out-of-vocabulary words. It also exhibits difficult challenges associated with code switching and regional dialects. This directly relates to an increase of difficulty for both ASR and SMT systems. Since SLT systems are generally built by feeding the output of the ASR system to an SMT system, each trained on separate datasets [1, 2], errors produced by the systems compound.

With respect to Egyptian Arabic specifically, unscripted, spontaneous, telephone conversations have been available through the Callhome Egyptian Arabic corpus (speech and transcripts) since 1997. However, since this dataset did not come with translations for the transcriptions in Arabic, researchers had to resort to using out-of-domain data to train the SMT systems. Transcripts for spontaneous conversations (speech), vary significantly from transcripts for scripted conversations and informal written conversations (web, forum, SMS, chat).

To bridge this gap between the type of data the ASR and SMT systems were trained on for SLT applications, we have created the Callhome Egyptian Arabic Speech Translation dataset. This supplements the existing LDC corpus with four reference translations for each utterance in the transcripts. The result is a three-way parallel dataset of Egyptian Arabic Speech recordings, transcriptions of the Arabic speech, and translations into English.

The primary goal of this paper is to describe the process of creation of this corpus in time for its pending public release, so that researchers who use the corpus have a good understanding of both its scope and limitations. We believe that this corpus will enable considerable new research in translation of spontaneous/conversational Arabic speech into English.

This work was supported by NSF IIS award No 0963898 and DARPA BOLT contract No HR0011-12-C-0015. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of NSF, DARPA or the U.S. Government.

LDC Catalog Number	Name	#train	#dev	#eval
LDC97S45, LDC97T19	Callhome Egyptian Arabic Speech/Transcripts	80	20	20
LDC2002S22, LDC2002T39	1997 HUB5 Arabic Evaluation	0	0	20
LDC2002S37, LDC2002T38	Callhome Egyptian Arabic Speech/Transcripts Supplement	0	0	20

Table 1. Sizes (in # conversations) of the Callhome Egyptian Arabic corpus, supplements and evaluation datasets. The conversations last between 5-30 minutes.

Partition	# Utt's	# Words	Words/Utt
ECA-96 (train)	20,861	139,035	6.66
ECA-96 (dev)	6,415	34,543	5.38
ECA-96 (test)	3,044	16,500	5.42
97-eval-H5	2,800	18,845	6.73
ECA-supplement	2772	18039	6.51

Table 2. Partition statistics for the Callhome Egyptian Arabic corpus, supplements and evaluation datasets. Column 2,3 and 4 represent number of utterances, numbers of words and average number of words per utterance respectively.

2. CORPUS AND TRANSLATION SETUP

We present English translations of the Egyptian-Arabic Callhome corpus, supplements and evaluation sets. These datasets were commissioned and used by the DARPA GALE (Global Autonomous Language Exploitation), DARPA EARS (Effective Affordable Reusable Speech-to-text) and the NIST HUB-5 LVSCR (Large Vocabulary Conversational Speech Recognition) programs.

The speech part of the corpus consists of unscripted telephone conversations between native speakers of Egyptian Colloquial Arabic (ECA). The conversations last between 5-30 minutes. In addition to the conversations, speaker metadata including gender, age, education and accent is available. Conversation metadata includes channel quality, crosstalk identifiers and number of speakers.

For each of the conversations, transcripts that cover a contiguous 5-10 minute segment are available. Manual audio segmentation information is available through the transcripts which have start and end time for each utterance in a conversation. Since ECA does not have a standard orthographic system, the conversations were transcribed using a romanized orthographic system which was phonemically based. This system preserves word pronunciation information and word identity. These transcripts in romanized orthography were then converted to Arabic script (encoding : ISO 8859-6) using a lexicon lookup [LDC99L22]. Table 2 provides details about this corpus.

2.1. Callhome Egyptian Arabic Speech/Transcripts

This corpus [Speech: LDC97S45, Transcripts: LDC97T19], hereafter referred to as ECA-96, consists of 120 unscripted telephone conversations. The corpus is split into three partitions : train, dev and eval, containing 80, 20 and 20 conversations respectively. The transcripts contain 30,320 utterances with a total of 190,078 words.

2.2. 1997 HUB5 Arabic Evaluation

This corpus [Speech: LDC2002S22, Transcripts: LDC2002T39], hereafter referred to as 97-eval-H5, was used as the evaluation set for the 1997 NIST HUB-5 non-English evaluation of conversational speech recognition systems. It consists of 20 unscripted telephone conversations. The transcripts contain 2,800 utterances with a total of 18,845 words.

2.3. Callhome Egyptian Arabic Speech/Transcripts Supplements

This corpus [Speech: LDC2002S37, Transcripts: LDC2002T38], hereafter referred to collectively as the ECA supplement, was initially sequestered for future NIST evaluations, but later released as a supplement to ECA-96. It consists of 20 unscripted telephone conversations. The transcripts contain 2,722 utterances comprised of 18,039 words.

2.4. Special Symbols in Transcription

Since the telephone conversations in this corpus are informal in nature and unscripted, special symbols are used to mark sections of the conversation that are not conventional Arabic speech. These contain non-verbal vocalizations, disfluencies, background noise and distortion. Table 3 provides a sample of some of these special symbols. Further details are available in the documentation of the respective corpus.

2.5. Egyptian Arabic Lexicon

An Egyptian Arabic colloquial pronunciation dictionary which supplements the corpora mentioned above, is available (LDC99L22). The lexicon contains 51,202 entries from the ECA-96, ECA-supplement and the Badawi and Hines dictionary of Egyptian Colloquial Arabic. This lexicon includes

Symbol	Interpretation
{text}	sound made by the talker
[text]	background or channel sound
<language text>	speech in another language
((text))	unintelligible, best guess provided
(( ))	unintelligible; can't guess text
text	idiosyncratic word, not in common use
-text , text-	partial words

Table 3. A sample of the special symbols using in the Arabic transcripts. These represent non-conventional speech segments such as non-verbal vocalizations, disfluencies, background noise and distortion.

orthographic representation of words in the LDC romanization scheme and Arabic script along with morphological, phonological, stress, source, and frequency information.

3. TRANSLATION METHODOLOGY

The translations for the Egyptian Arabic Callhome corpus were obtained using crowd-sourcing techniques. Crowdsourcing has become a standard technique in the collection and annotation of scientific data [3, 4, 5, 6, 7, 8] including data for natural language processing tasks like machine translation [9]. We use the crowdsourcing platform, Amazon Mechanical Turk (MTurk) to obtain translations. We follow the best practices suggested by [9] in this process.

3.1. Pre-processing

Each transcript was pre-processed to remove markup, including the special symbols described in Section 2.4. Some special symbols contain text in a foreign language (mostly, English). These were retained so that they could be passed through to the translation. Utterances that comprised only of markup and the special symbols were removed. Each utterance in the corpus contains channel and segment information. These were incorporated as a part of a segment identifier so that the translations could be mapped back to the transcriptions and the speech segments.

3.2. Collecting Translations

A translation task on the MTurk platform is presented to the translators as a HIT (Human Intelligence Task). Each translator was presented with a sequence of ten segments to translate. These segments or utterances were always presented in the order they appear on the transcripts. Since the conversation consists of two channels, the order presented generally comprised of alternating speakers. This allowed the translators to incorporate context wherever it is available and helpful. Each HIT included translation instructions derived from [9]. In addition, translators were instructed to retain the foreign language information in the utterance. As noted earlier, the transcripts were converted to Arabic script from an intermediate romanized version. We did not attempt to normalize any non-MSA words to create an MSA equivalent. Four independent translations were obtained for each utterance using such HITs.

3.3. Quality Control

MTurk provides a quality control mechanism which relies on vetting of users and qualification tests. However, these methods in isolation are not enough to guarantee high quality translations. We used the following quality control measures to ensure that the quality of the translations was acceptable and to prevent inappropriate use of the platform.

- For each utterance, we obtained translations from Google Translate. If a translation had a small edit distance from the translation obtained via Google Translate, it was flagged, reviewed and rejected if it had the same errors.
- The utterance for translation task was presented as an image rather than as text. This prevented users from using online translation services to cut and paste translations.
- Manually translated gold standard segments were inserted into our dataset. Each translator was presented with three such segments. Their HITs were flagged, reviewed and rejected if their translation for these segments was not similar to the gold standard translations.
- We gathered self-reported geographical and language information for each of our contributors on MTurk. The specification for our task asked native Arabic speakers to participate. Since HITs had to be manually approved, we checked translator metadata and number of translations received. In addition, prior to approval, a spot check of the translations was conducted. Finally, higher preference was given to trusted Arabic speakers that we have worked with on other translation tasks.

3.4. Post-processing

The translations were split based on the partitions described in Section 2 and each partition was duplicated (typically four-fold) to obtain redundant/independent translations. For some utterances, we ended up obtained more than four translations. These were stored in an overflow file. Utterances that only contained markup and special symbols (which were previously removed) were re-inserted into this set of translations to restore utterance-level synchronization with the LDC corpora.

Partition	# Utt's	# Words	Words/Utt
ECA-96 (train)	86,313	713,549	8.27
ECA-96 (dev)	25,769	186,400	7.23
ECA-96 (test)	12,212	85,182	6.98
97-eval-H5	11,248	91,647	8.15
ECA-supplement	11,126	87,489	7.86

Table 4. The results of the translation task described in section 4. Each utterance in the original partitions has about four redundant translations. The number of utterances in column 2 has hence effectively been multiplied by 4. The last column represents the number of words per utterance in the translations.

Partition	Crossfold BLEU
ECA-96 (train)	40.09%
ECA-96 (dev)	35.64%
ECA-96 (test)	35.86%
97-eval-H5	35.81%
ECA-supplement	37.15%

Table 6. Inter-annotator BLEU per partition of the Callhome Egyptian Arabic corpus, supplements and evaluation datasets. Each translation was evaluated against three translations to obtain a BLEU score per utterance. This was averaged per partition.

4. TRANSLATION TASK RESULTS AND CONSISTENCY

In total, 838 translators participated in this process, producing 143,568 translations in English. Table 4 summarizes the results of the translation task. Note that the average number of words per utterance has increased after translation to English. Table 5 provides a sample of the translations obtained.

To measure inter-annotator agreement, we used a cross-folding type BLEU scoring scheme. Translations for each partition were lower-cased, tokenized using the Penn WSJ treebank conventions, and punctuation was normalized. Each translation was then evaluated against the remaining three translations in a cross-folding fashion. The results were averaged per dataset partition. The results of these experiments are in Table 6.

5. PLANNED CORPUS RELEASE

In a manner similar to the previous work on speech translation of [10], based on the Spanish Fisher and Callhome corpora, we plan to provide Automatic Speech Recognition (ASR) output for the datasets in the Callhome Egyptian Arabic corpus. The ASR output will be provided in the form of OpenFST lattices, lattice oracles (paths that have the least word error rate in the lattice) and the 1-best output. This will effectively lead to the creation of a four-way parallel dataset with Egyptian Arabic speech, transcripts, ASR output and English translations. Our goal in providing the ASR output is to enable research in speech translation for Statistical Machine Translation (SMT) researchers as well.

6. CONCLUSION

We presented the Callhome Egyptian Arabic Speech Translation Corpus based on the Callhome Egyptian Arabic corpus, supplements and evaluation (HUB5) datasets. With the ASR output, the resulting speech translation corpus is a four-way parallel dataset with Egyptian Arabic speech, transcripts, ASR output (lattice, lattice oracle and 1-best) and translations. This in-domain dataset is an effort to aid research in translation of spontaneous, conversational speech with a long term goal of improving human-human conversation.

7. REFERENCES

[1] Oliver Bender Richard Zens, "The RWTH phrase-based statistical machine translation system," in *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, 2005, pp. 155–162.

[2] E. Matusov, S. Kanthak, and H. Ney, "Integrating speech recognition and machine translation: Where do we stand?," in *2006 IEEE International Conference on Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings*, May 2006, vol. 5, pp. V–V.

[3] Scott Novotney and Chris Callison-Burch, "Cheap, fast and good enough: Automatic speech recognition with non-expert transcription," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Stroudsburg, PA, USA, 2010, HLT '10, p. 207215, Association for Computational Linguistics.

[4] A Sorokin and D. Forsyth, "Utility data annotation with amazon mechanical turk," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2008. CVPRW '08*, June 2008, pp. 1–8.

[5] Aniket Kittur, Ed H. Chi, and Bongwon Suh, "Crowdsourcing user studies with mechanical turk," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2008, CHI '08, p. 453456, ACM.

[6] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng, "Cheap and fastbut is it good?: Evaluating non-expert annotations for natural language tasks," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA,

Source	ما انتو مبتردوش على التليفون يبقى
Translation 1	you do n't reply to the phone
Translation 2	so you do n't answer the phone then
Translation 3	you do n't answer the phone it seems
Translation 4	because you do n't answer the call then
Source	مصعبان عليه نفسه كمان
Translation 1	he feels hard for himself too
Translation 2	he feel bad about himself
Translation 3	he feels sorry for himself too
Translation 4	i feel sorrow about his condition too

Table 5. A sample of the translations obtained using the translation task described in section 4. The translations are lower-cased, tokenized and punctuation has been normalized.

USA, 2008, EMNLP '08, p. 254263, Association for Computational Linguistics.

[7] Chris Callison-Burch, "Fast, cheap, and creative: Evaluating translation quality using amazon's mechanical turk," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, Stroudsburg, PA, USA, 2009, EMNLP '09, p. 286295, Association for Computational Linguistics.

[8] Gabriele Paolacci, Jesse Chandler, and Panagiotis G. Ipeirotis, "Running experiments on amazon mechanical turk," SSRN Scholarly Paper ID 1626226, Social Science Research Network, Rochester, NY, June 2010.

[9] Omar F. Zaidan and Chris Callison-Burch, "Crowdsourcing translation: Professional quality from non-professionals," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, Stroudsburg, PA, USA, 2011, HLT '11, p. 12201229, Association for Computational Linguistics.

[10] Matt Post, Gaurav Kumar, Adam Lopez, Damianos Karakos, Chris Callison-Burch, and Sanjeev Khudanpur, "Improved speech-to-text translation with the fisher and callhome translated corpus of spanish-english speech," in *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, 2013.