Translation of Egyptian-Arabic Conversational Telephone Speech

Gaurav Kumar

Advisors : Graeme Blackwood, Yaser Al-onaizan, Abe Ittycheriah
IBM Research, Johns Hopkins University
gkumar@cs.jhu.edu

September 11, 2014

Overview

1. Overview of the problem

2. Interfacing ASR and SMT

3. SMT strategies and results

4. Conclusion

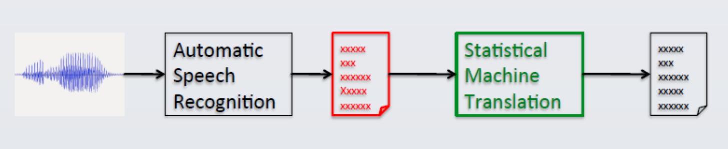
Speech Translation

Speech Translation

To translate speech in one language (eg. Arabic) to text/speech in another language (Eg. English)

This generally involves :

- Get automatic transcripts from speech (ASR)
- Get translations of these automatic transcripts (SMT)
- Convert translations to speech (TTS)



The Egyptian Arabic Callhome Corpus

- Callhome Egyptian Arabic Speech/Transcripts (ECA-96 : train, dev, test)
- 1997 HUB5 Arabic Evaluation (97-eval-H5)
- Callhome Egyptian Arabic Speech/Transcripts Supplement (ECA-supplement)

Partition	**# Conv**	**# Utt's**	**# Words**	**Words/Utt**
ECA-96 (train)	80	20,861	139,035	6.66
ECA-96 (dev)	20	6,415	34,543	5.38
ECA-96 (test)	20	3,044	16,500	5.42
97-eval-H5	20	2,800	18,845	6.73
ECA-supplement	20	2772	18039	6.51

Table: Partition statistics for the Callhome Egyptian Arabic corpus, supplements and evaluation datasets.

The Egyptian Arabic Callhome Corpus : Translations

We currently use translations of the Callhome Egyptian Arabic Corpus generated at JHU [1]. **Four** alternative translations were created for each utterance in the corpus.

- Collected via Amazon's Mechanical Turk platform.
- Translations by native speakers of the source language.
- Quality control and post editing done.

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: Reference (English) translation statistics for the Egyptian-Arabic Callhome corpus.

[1] Kumar et al., *Speech Translation with the Callhome Egyptian Arabic Corpus*, IEEE SLT, 2014 (submitted)

The Egyptian Arabic Callhome Corpus : Translations

- Inter-translator agreement is low.
- This is typical for human translations of informal speech/text.
- Diversity in translations required to compensate for possibility of multiple valid answers automatic SMT evaluation.

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: Crossfold (hold-1, against-3) average BLEU per partition of the Callhome Egyptian Arabic corpus, supplements and evaluation datasets.

The Egyptian Arabic Callhome Corpus : Translations

Source	mA Antw mbtrdw$ ElY Altlyfwn ybqY
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	mSEbAn Elyh nfsh kmAn
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: A sample of the translations for the Egyptian-Arabic Callhome Corpus. The translations are lower-cased, tokenized and punctuation has been normalized.

The ASR system

- The ASR system is built using Kaldi @ JHU
- Training data for acoustic modeling : $\sim$ 16 hours
- Automated evaluation using the NIST sclite scoring tool
- This is work in progress, more data ($\sim$ 30 hours) recently available through LDC (+ $\sim$ 100 hours from GALE)

ASR Model	WER
Triphone + SAT (train)	57.98%
+ SGMM (dev)	52.74%
+ DNN (test)	51.80%

Table: Word error rates (WER) for ASR models

Results in this presentation use the output of the Triphone + SAT ASR system.

Decoder modes and their efficacy for Speech Translation

Our work uses the IBM Egyptian-Arabic to English machine translation system. Three options in SMT models for decoding were available to us :

- Monotone Phrase based models (Monotone)
- Hierarchical Phrase-based models (Hiero)
- Tree to String models (T2S)

SMT model	**(T-B)/2)**
T2S	15.66
Hieros	15.98
Monotone	17.61

Table: (T-B)/2 scores for DF-dev with different decoder modes. T2S provides a very small gain over Hieros. Hieros provide a significant gain over Monotone models.

The effect of punctuation

ASR output does not typically contain punctuation. The phrase table built during SMT training does contain punctuation. Removing puncutation from the phrase table :

- Remove punctuation from the *source* side in the phrase tables. Target side punctuation is retained.
- Merge counts for duplicates that arise
- Re-calculate model 1 scores
- **No punctuation in the source represents the evaluation condition for CTS translation.**

Source Punct ?	**Grammar Punct ?**	**(T-B)/2**
+	+	19.02
-	-	**18.06**
+	-	19.20
-	+	**18.22**

Table: (T-B)/2 scores for Monotone decoding CTS-REF-TRAIN40 with and without punctuation in the source and grammar

Selecting appropriate ASR output

ASR 1-best output

The best hypothesis based on Acoustic model and Language models scores for an utterance.

ASR-lattice (L)

A compact graph representation (typically encoded as a weighted finite state acceptor) of all of (typically pruned) valid hypotheses for an utterance. Weights on edges are an interpolation on the acoustic model (AM) and language model (LM) scores.

ASR-oracle

The hypothesis corresponding to a path in the ASR-lattice that has the least WER.

Selecting appropriate ASR output

What is the right interface between ASR and SMT systems ?

- ASR 1-best : Rely on the ASR system to choose the best hypothesis for SMT.
- ASR word-lattice : Use lattice decoding to choose the best translation from among multiple hypotheses.

The ASR system has no infomation about the purpose that its output will be used for. How can we incorporate this ?

	ASR 1-best	**ASR Oracle**
WER	57.98%	33.76%
Insertions	1,890	2,020
Deletions	4,397	2,082
Substituions	13,389	6977

Table: ASR performance for 1-best vs. Oracle for the Triphone+SAT ASR system for ARZ-CTS.

Selecting appropriate ASR output

We present three strategies for selecting better output for translation from the ASR lattice. They utilize the following components:

Phrase segmentation Transducer (S)

A finite state transducer (FST) that is built from the SMT phrase table. This machine transduces a sequence of words to a sequence of phrases. Each entry in the phrase table corresponds to a path (cyclic) in this FST.

Phrase lattice (P)

The result of the composition of a word lattice (acceptor) with the phrase segmentation transducer. This represents all possible phrase segmentations of the hypotheses in the word lattice.

Generating the phrase lattice

$$P = \det(\min(L \circ S))$$

Maximum Spanning Phrases : A toy example

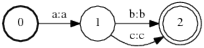


Figure: A word lattice.

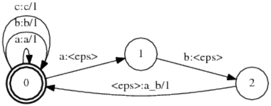


Figure: A segmentation transducer

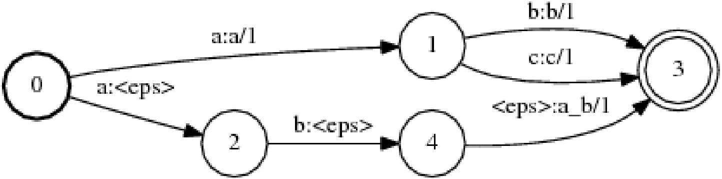


Figure: A phrase lattice produced by composing the word lattice with the segmentation transducer.

Maximum spanning phrases

Strategy 1 : MSP-greedy (WER = 60.0% : -2.02%)

Select the hypothesis that needs the least number of phrases to cover it (The shortest path in the phrase lattice). The word lattice and the phrase segmentation transducer are unweighted in this strategy.

Strategy 2 : MSP-ASR-draws

In MSP-greedy if there are draws in length, select the hypothesis that has the best ASR score among the draws. The word lattice and the phrase segmentation transducer are unweighted in this strategy.

Strategy 3 : MSP-Unigram

Create the phrase lattice (P) by composing weighted ASR lattices with the segmentation transducer(S). Paths in P are weighted using the phrase unigram probability.

$$w_(\pi_i) = \bigotimes_{p_j \in project(\pi_i)} cost(p_j) \;\text{ where }\; cost(p_j) = \frac{freq(p_j)}{\sum_k freq(p_k)}$$

where π is a path in P and p_i is a phrase in the phrase table.

Strategy 4 : MSP-Unigram-Length

To avoid penalizing phrases that are longer and appear less often, scores for paths in S are normalized by the count of phrases with the same length.

$$w_(\pi_i) = \bigotimes_{p_j \in project(\pi_i)} cost(p_j) \;\text{ where }\; cost(p_j) = \frac{freq(p_j)}{\sum_{k\ st\ len(p_k)=len(p_j)} freq(p_k)}$$

where π is a path in P and p_i is a phrase in the phrase table..

Discussion Forums (DF) decoder baseline results

Decoding setup :

- Evaluated set : ECA-96 (dev) set.
- ASR ouput from Triphone + SAT system.
- No puncutation in the source and grammar
- Monotone phrase based model used.
- All experiments are replicated for seven input types : Human transcripts, ASR 1-best, ASR-Oracle, MSP-Greedy, ASR-draws, Unigram, Unigram-length

Input type	**R2S**	**(T-B)/2**
REF	0.86	25.49
ASR 1-best	0.90	39.51
ASR-Oracle	0.89	33.50

Table: Baseline CTS decoding results with the DF SMT system.

Input type	**R2S**	**(T-B)/2**
MSP-greedy	0.97	35.80
MSP-ASR-Draws	0.95	36.86
MSP-UNI	0.96	36.56
MSP-UNI-LEN	0.94	37.36

Table: Baseline CTS decoding results with the DF SMT system for MSP input.

Tuning on CTS data

- Tuning of the decoding parameters was done on the 97-EVAL-H5 dataset.
- The hope was to adjust for length (R2S) in the baseline experiment.
- Tuning boosted translation model weights and reduced the weights for lm and wc.
- Reduction in insertion errors (13667 vs 6186) leads to better decoding (mainly TER) results.

	Baseline		**+Tune-REF**	
Input type	**R2S**	**(T-B)/2**	**R2S**	**(T-B)/2**
REF	0.86	25.49	0.97	16.42
ASR 1-best	0.90	39.51	1.05	30.75
ASR-Oracle	0.89	33.50	1.03	24.64
MSP-greedy	0.97	35.80	1.16	29.49

Table: CTS decoding results with tuning on CTS data.

Tuning the LM interpolation weights

- Tuning revealed that a more appropriate interpolation of the weights for the LM components may be needed.
- Perplexity tuning was done for the LM component interpolation weights using the ECA-supplement dataset
- Weights increased for the BOLT LM and reduced for everything else.

	+Tune-REF		+Tune-REF+LM(exp)	
Input type	**R2S**	**(T-B)/2**	**R2S**	**(T-B)/2**
REF	0.97	16.42	0.97	16.35
ASR 1-best	1.05	30.75	1.05	30.65
ASR-Oracle	1.03	24.64	1.02	24.52
MSP-greedy	1.16	29.49	1.16	29.39

Table: CTS decoding results, tuned on CTS data, tuned LM weights

Tuning on ASR output

- Tuned on ASR 1-best output for 97-H5-eval.
- Decoding yields best R2S but worse ($\sim$ points) (T-B)/2 results.
- Tuning set significantly smaller because of the large number of empty sentences.
- Tuning on high WER may cause this, results may change with lower WER ASR output.

	+Tune-REF+LM(exp)		**+Tune-ASR+LM(exp)**	
Input type	**R2S**	**(T-B)/2**	**R2S**	**(T-B)/2**
REF	0.96	17.36	0.95	18.17
ASR 1-best	1.03	31.37	1.02	31.96
ASR-Oracle	1.00	25.17	0.99	25.86
MSP-greedy	1.13	29.76	1.11	30.00

Table: CTS decoding results, tuned on non-empty ASR output, tuned LM weights

A hope for the future : Decoding ASR output with lower WERs

Decoding ASR output with a lower WER results in a significant gain in (T-B)/2 scores.

ASR-model	WER	R2S	(T-B)/2
Triphone+SAT	57.98%	0.95	30.65
+SGMM 1-best	52.74%	0.95	28.55
+DNN	51.80%	0.96	28.58

Table: CTS decoding results for the ASR 1-best from three difference ASR models with lower WER. The SMT system parameters and the LM weights are tuned on CTS data

Backchanneling

- The ASR system frequently misrecognizes very short utterances with hesitations/non-vocal markers.
- These mis-recognitions are typically utterances produced by back-channeling.
- Hence the ASR output after stripping these is much shorter (1027 additional empty segments)

Utterance	Frequency
yes	476
m	131
mm	72
what	39
yes yes	29
ha	19
ok	18
umm	13
na	8
aha	7
m m	6
yes yes yes	4

Table: Frequent utterances misrecognized by the ASR for hesitation/non-vocal sounds

Conclusion

Future work

- Validate results on ASR output with lower WER.
- Tuning on ASR output and human transcripts.
- Include CTS data in the LM for SMT
- **Most common misrecognitions from the ASR can be added to the SMT phrase table. This should help correct these common mistakes.**
- **Apply the concept of selecting hypotheses from the lattice (MSP) for Hieros.**

Questions?