Unsupervised Arabic Dialect Adaptation with Self-Training

Scott Novotney[1], *Rich Schwartz*[1], *Sanjeev Khudanpur*[2]

[1] BBN Technologies, Cambridge, MA, USA
[2] HLTCOE and ECE Dept., Johns Hopkins University, Baltimore, MD, USA

snovotne@bbn.com, schwartz@bbn.com, khudanpur@jhu.edu

Abstract

Useful training data for automatic speech recognition systems of colloquial speech is usually limited to expensive in-domain transcription. Broadcast news is an appealing source of easily available data to bootstrap into a new dialect. However, some languages, like Arabic, have deep linguistic differences resulting in poor cross domain performance. If no in-domain transcripts are available, but a large amount of in-domain audio is, self-training may be a suitable technique to bootstrap into the domain. In this work, we attempt to adapt Modern Standard Arabic (MSA) models to Levantine Arabic without any in-domain manual transcription. We contrast with varying amounts of in-domain transcription and show that 1) Self-training is effective with only one hour of in-domain transcripts. 2) Self-training is not a suitable solution to improve strong MSA models on Levantine. 3) Two metrics that quantify model bias predict self-training success. 4) Model bias explains the failure of self-training to adapt across strong domain mismatch.

Index Terms: Arabic ASR, domain adaptation, self-training

1. Introduction

Effective large vocabulary continuous speech recognition (LVCSR) of spoken colloquial dialects requires costly in-domain transcription since these languages are not formally written. For example, given the dozens of Hindi sub-dialects, deploying a system with low word error rate (WER) (trained on 100 hours of transcription) for each one is infeasible. However, for almost any language, broadcast news (BN) corpora are much more easily available as well as copious amounts of news wire and web data for language modeling.

In this paper, we focus on LVCSR of conversational telephone speech (CTS) for spoken colloquial dialects. Unfortunately, some languages have wider domain mismatches between BN and CTS. Dialectal Arabic is one example of this mismatch. Systems trained on broadcast news, prompted speech or other colloquial dialects show very poor cross-domain performance on CTS data. Without in-domain transcription, typical domain adaptation techniques appear useless.

Previous work [1] drastically reduced WER of poor initial acoustic models by decoding large amounts of data and training on the automatic transcripts, despite high initial WER. This method, self-training (or unsupervised training), was used to successfully adapt Spanish BN acoustic models (AM) to Polish BN, reducing WER from 63.4% to 20%. Even well-trained models can be improved. Self-training of 150 hours of MSA reduced WER from 16.7% to 15.5% [2].

Most previous work on self-training used strong language models (LM) since the target domain was BN with lots of available newswire data. Work with English CTS acoustic modeling showed reductions in WER from 58% to 37% with only 100k words of in-domain transcriptions [3]. Our work differs from the previous literature in that *no* in-domain transcription or language modeling text will be available. We will also compare performance with stronger in-domain LMs.

Work on colloquial Arabic adaptation from MSA has focused on mapping between phonemes and then using techniques like MAP and MLLR to improve system performance. Adapting 33 hours of MSA with 20 hours of transcribed Egyptian reduced WER by 6% [4]. A reduction in WER from 16.8% to 11.8% was achieved by first adapting 12 hours of MSA to 45 hours of phonetically transcribed Tunisian prompted telephone speech and then using MLLR to adapt to the final Jordanian test data [5]. We hope a large amount of unlabeled audio (100 hours) will compensate for a smaller amount of transcribed data. Although we are working with Arabic, our goal is not to improve state-of-the-art dialectal LVCSR performance. Instead, we use this as a test bed for unsupervised adaptation across domain and dialect.

In this paper, we adapt a strong MSA acoustic and language model to Levantine Arabic using only a list of Levantine words and 100 hours of unlabeled audio. Self-training must compensate for differences in acoustic channel, speaking style and language. Sections 2 and 3 describe the system and corpora and quantify the differences between MSA and Levantine. Section 4 provides supervised baselines and upper bounds for unsupervised adaptation. Section 5 shows that in-domain self-training with one hour of Levantine manual transcripts behaves similar to previous work, giving significant gains. Section 6 then attempts to adapt from MSA to Levantine using both self-training and standard adaptation techniques of MLLR and MAP but with limited gains. Finally, section 7 explains why in-domain transcription succeeded while adaptation failed and offers a strong predictor of self-training effectiveness.

2. System Description

We used a multi-pass state-of-the-art LVCSR system that uses state-clustered Gaussian tied-mixture models [1]. Decoding requires three passes: a forward and backward pass using triphone models and an approximate trigram LM to generate an N-best list, which is then rescored using quinphone cross-word acoustic models and trigram LM. These three steps are then repeated after speaker adaptation using constrained maximum likelihood regression. We do not run discriminative training since little gain has been shown for unsupervised scenarios [7].

For self-training, we decode a large amount of unlabeled audio and for each utterance estimate the confidence that the WER is below some threshold. Then we retrain and iterate until WER stops decreasing, typically two or three passes. While finer-grained selection at the word or frame level would help, the gain versus utterance level selection is not large [8]. The original MSA transcripts are not included in the new acoustic models. We run AM and LM self-training independently, leaving the other model fixed, in order to control for WER reductions.

3. Corpora

In order to compare our semi-supervised techniques with fully supervised transcription, we require that our target corpus be manually transcribed. Additionally, since self-training works best with large amounts of audio [3], we selected the Levantine Arabic Fisher corpus. We refer the reader to previous literature [8], for detailed explanations of the linguistic differences between dialects of Arabic.

3.1. Levantine Arabic

Spoken by 35 million people in the Levant region – Syria, Palestine, Jordan, and Lebanon – this dialect lacks case inflection and is lax on gender/number agreement. We use 156 hours in total from LDC corpus LDC2007T04. This corpus follows the Fisher transcription methodology, where strangers are assigned to speak about one of a few dozen topics for ten minutes over cellular or landline telephones.

The 156 hours were partitioned into one or ten hour initial training sets, 100 hours of "unlabeled" audio for decoding, and the remaining 45 hours used to build a strong in-domain LM. Three LMs were built from 1, 10 or 45 hours of language modeling text and the 100 hours were held out for use in decoding during self-training. For testing, we used 2.5 hours of carefully transcribed data released as part of the RT 2004 Arabic CTS evaluation, again without short vowels. We used a 37k graphemic dictionary derived from the released LDC corpora for all experiments. This is the only supervised resource assumed available for adaptation.

3.2. Modern Standard Arabic

The literary standard of the Arab world, MSA is extensively used in broadcast news and newswire. It is not, however, a native dialect used in conversation. The MSA system was trained with 1400 hours of transcribed BN. This system when used on MSA BN test data gives 11% WER. For decoding Levantine, the audio was band limited to 8 kHz to match the Levantine telephony data and the word models were derived from the 37k Levantine dictionary. Since both systems used graphemic pronunciations, the phoneme set was the same.

3.3. Differences between MSA and Levantine

Besides acoustic differences, the vocabularies of the two languages are very different. The 256k MSA vocabulary gave a 29% out of vocabulary (OOV) rate on the Levantine test set. For comparison, the 37k Levantine vocabulary has an OOV rate of 6%. Since we assume this dictionary is given to us, this is the OOV rate of all models in this work. Morpheme-based decoders would help, but the complexity of tackling this issue has not shown a significant gain [10].

Additionally, Levantine is primarily SVO word ordering while MSA is predominantly VSO. As seen in Table 1, this leads to MSA being a very poor language model for Levantine. 10M words of MSA are significantly worse than 7K words of Levantine transcripts. For comparison, 1M words of English BN are about as strong as 200K words of English Fisher CTS transcripts. MSA and Levantine are not just separate dialects, but indeed, different languages.

If we instead ask whether the two words in a Levantine bigram appear *anywhere* in the MSA text (not just in sequence) the "component" hit rates double, hinting at the VSO/SVO mismatch. Devising a mapping between MSA and Levantine n-grams completely unsupervised does not appear feasible without extensive linguist knowledge of the two domains.

		N-gram Count		Ngram Hit Rate	
LM	**PPL**	**2gr**	**3gr**	**2gr**	**3gr**
1400hr MSA	3830	55M	45M	6%	1%
1hr Lev	1220	6k	7k	25%	5%
10hr Lev	709	45k	60k	42%	12%
45hr Lev	521	150k	245k	54%	19%

Table 1 – *Comparison of LM strength on Levantine.* Four different language models (rows 2-5) were evaluated against Levantine. The perplexity (col. 2) encapsulates the power of the LM. N-gram counts (col. 3-4) show the number of unique n-gram types. Hit rate (col. 5-6) measures the percentage of test n-grams (by token) that appeared in the LM training data. Despite millions of n-grams, the MSA LM is much poorer than one hour of in-domain Levantine. These four LMs will be paired with acoustic models for self-training experiments.

4. Supervised Baselines

We first measured WER on Levantine test data using different amounts of manual transcription for AM and LM training. This defines the landscape for gains in self-training.

		Language Model				
		MSA	*Levantine*			
			1hr	*10hr*	*45hr*	*100hr*
Acoustic Model	*1400 MSA*	**69.8**	68.8	63.9	61.4	**61.1**
	1hr Lev	84.8	79.0	76.7	75.2	75.1
	10hr Lev	70.1	65.2	62.5	60.1	59.5
	100hr Lev	**59.1**	55.0	52.9	50.5	**50.1**

Table 2 – *Upper bounds for self-training.* Four different AMs (rows) were paired with five different LMs (columns). The 1400hr MSA starting point (top left cell) has an initial WER of 69.8%. 100 hours of Levantine manual transcription would improve the acoustic model (bottom left) and reduce WER by 11%. There is a 9% gain for the LM (top right). Combining these two results gives the total possible gain for 100 hours of Levantine manual transcription (bottom right).

Notice from Table 2 that the 1400hr MSA acoustic model has similar WER to ten hours of Levantine as language models change in strength. Similarly, as seen in Table 3, ten hours of MSA has about the same strength as one hour of Levantine data for acoustic modeling. Additional MSA acoustic training very slowly reduces WER on Levantine. For the language model, the situation is even worse. As seen in Table 1 and Table 2, the MSA LM is significantly worse than one hour of Levantine in terms of both perplexity and WER when paired with either MSA or Levantine acoustic models.

		Language Model	
		MSA	*45hr Lev*
Acoustic Model	*1 Lev*	84.8	75.2
	10 MSA	82.5	75.3
	10 Lev	70.1	60.9
	1400 MSA	69.8	61.4

Table 3 – *Equivalent MSA and Levantine models.* When paired with two different LMs (columns), one hour of Levantine and 10 hours of MSA have about the same strength (WER) when used as an AM. Similarly, 10 hours of Levantine and 1400 hours of MSA are of the same strength. BN MSA is a poor acoustic match to CTS Levantine.

The goal of adaptation is to decrease the starting WER of 70% to as close to 50% as possible by adapting both the acoustic and language model. We decoded 100 hours of Levantine data and retrained new acoustic and language models from these highly inaccurate automatic transcripts. In addition to having only the MSA data available, we also experimented with different strength LMs quantifying the value of in-domain transcription or better LM techniques. Of course, if we had ten hours of data to use for language modeling, we would build an acoustic model as well. But we instead wanted to simulate the power of a language model of similar strength as 10 or 45 hours of in-domain transcripts.

To measure the success of self-training, we use *WER Recovery*. Given the WER of the initial model (I), the self-trained model (U) and the supervised model, (S), we measure the fraction of the gain from supervised training *recovered* by self-training. A recovery of 100% implies that self-training is as effective as manual training. A negative recovery means that the self-trained model is worse than the initial model.

$$WER\ Recovery = \frac{WER_I - WER_U}{WER_I - WER_S} \qquad (1)$$

5. Levantine AM Self-Training

We used one hour of Levantine to build an acoustic model and then three different LMs of varying strengths. Even with a weak LM trained on one hour of speech (7K words), we still reduced WER by 8% and achieve 32% of the total possible gain for improving the acoustic model. Even very weak LMs can be effective for self-training. Stronger LMs improved both WER and recovery, achieving 53% of the gain for manual transcription, similar to previous results [3].

AM	LM	Initial	Unsup.	Sup.	Recov.
1hr	*1hr*	79.0	71.3	55.0	32%
1hr	*10hr*	76.9	66.7	52.9	42%
1hr	*45hr*	75.2	61.9	50.5	53%

Table 4 – *Levantine Self-Training Results*. Using one hour of manual Levantine transcripts and three different LMs (rows 2-4) we decoded 100 hours of unlabeled data. The initial WER (col. 3) was improved after two iterations of self-training (col. 4). The gain for self-training was compared to the gain for manual transcription of the audio (col. 5) to give the WER Recovery (col. 6). We see WER reductions of 8%+ even with a poor one hour LM, but stronger LMs both reduced the initial WER and improve the effectiveness of AM self-training.

6. Adapting from MSA to Levantine

Since we assumed a vocabulary list of Levantine was given to us and used graphemic pronunciations, the two unsupervised tasks were adapting the MSA language and acoustic models. We used the best MSA system, with 1400 hours of transcripts, to decode our data. We also experimented with weaker MSA systems of 10 and 100 hours to gauge success at different operating points. So as not to conflate results, we only improve the LM or the AM, leaving the other model fixed during self-training.

6.1. Language Model Self-Training

Previous works showed that building an LM on automatic transcripts can improve performance, but is very sensitive to the WER [3]. We decoded the 100 hours of Levantine audio with the MSA models and built an LM using n-gram counts weighted by the product of individual word confidences. Despite only 12% of the bigram tokens being correct and 6% of trigrams, interpolating this LM with the 1400hr MSA LM reduced WER by 2% from 69.8% to 67.9%. If we instead had manual transcripts for the 100 hours, the WER would be 61%, resulting in a WER Recovery of (69.8 – 67.9) / (69.8 – 61.1) = 21%. This low recovery is in line with the previous LM self-training results [3], but still encouraging. Self-training increased the likelihood of the common Levantine filler words (*e.g. yeah*, *uh-huh*) not present in BN transcripts.

Attempts at extracting Levantine LM text in the MSA transcripts were unsuccessful. We ranked each MSA utterance by the OOV rate using the Levantine vocabulary and built LMs at various thresholds. None of these outperformed training on all the MSA data. As a final oracle test, we extracted from the MSA only those n-grams (up to trigrams) which appeared in the 100 hours of manual Levantine transcription. This *still* had higher perplexity than using all the MSA data. Even though each extracted n-gram was uttered by a Levantine speaker, the relative frequencies of these n-grams are poor estimates of Levantine speech.

6.2. Acoustic Model Self-Training

We first tried standard adaptation techniques of MAP and MLLR but saw little success. Adapting 1400 hours of MSA with 100 hours of automatically decoded Levantine (with 70% WER) reduced WER by only 0.1%. MAP showed no gains. While previous work showed great gains with manual transcripts, the automatic transcripts were of little value.

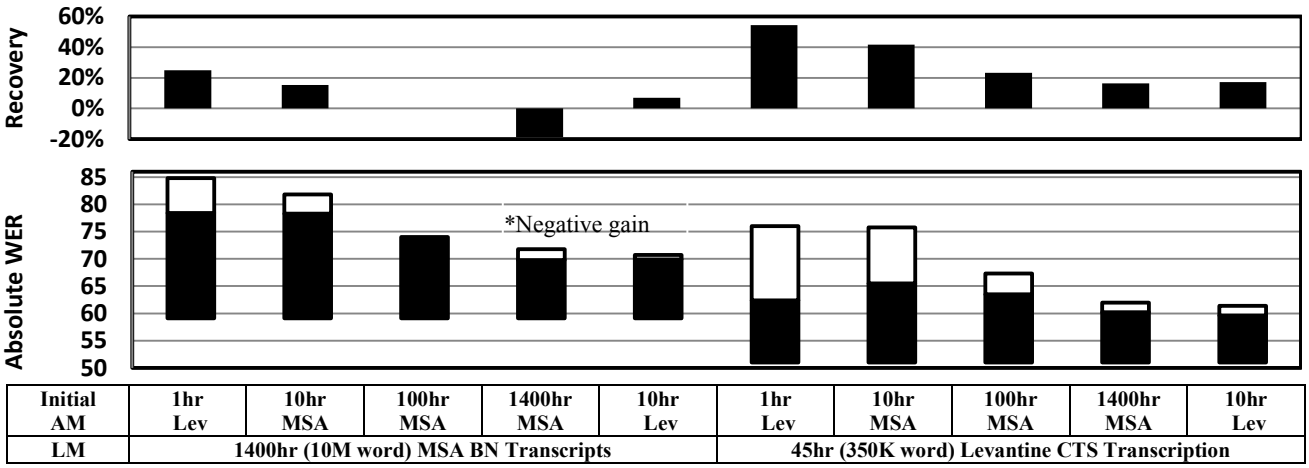
We then ran self-training on three different MSA AMs - 10, 100, and 1400 hours. These different strength MSA AMs reflect different operating points of available BN transcription, as not all languages will have 1000+ hours of acoustic transcription. For comparison, we also included one and ten hour Levantine AMs as we expect the one hour to have the highest WER Recovery while ten will be fairly low.

We paired these five AMs with two different language models: a fair LM trained on 1400 hours of MSA transcripts and an LM trained on 45 hours of Levantine manual transcripts. This reflects the best case scenario for self-training, as in most previous work. All experiments decoded 100 hours of Levantine audio held out from any supervised data. Additionally, all experiments used the same vocabulary.

Figure 1 details the experimental results for each of the ten conditions. Self-training with the best MSA system (col. 4) does not improve WER and in fact increases WER by 2%. The two weaker MSA models show small gains as well. One explanation for these negative results is the relatively small amount of 100 hours of Levantine.

During self-training, the original MSA data were not included when training new acoustic models. The MSA acoustic features would overwhelm the Levantine data. So the self-trained model only had 100 hours of training data, which could not improve over the 1400 hour baseline (col. 4). It did, however, improve over the 10 hour MSA starting condition (col. 2). More unlabeled audio would reduce the self-trained WER and give positive WER Recovery.

The negative results are not solely due to the MSA LM as it does have some value for AM self-training. The one hour Levantine and ten hour MSA models have positive, but small, recovery (cols. 1 and 2). With a strong in-domain LM, results were much more promising. All models reduced in WER and in this case, the 100 hours of Levantine outperforms the 1400 hour MSA AM (col 9). Not only does a strong in-domain LM reduce initial WER, but also improves the effectiveness of self-training by compensating for errors made by the acoustic model.



Initial AM	1hr Lev	10hr MSA	100hr MSA	1400hr MSA	10hr Lev	1hr Lev	10hr MSA	100hr MSA	1400hr MSA	10hr Lev
LM	1400hr (10M word) MSA BN Transcripts					45hr (350K word) Levantine CTS Transcription				

Figure 1 – *Self-Training Results.* Each column is a separate experiment consisting of an initial AM and LM detailed at the bottom. There are three WER values displayed in the middle section. The top of each bar is the WER of the initial AM and LM. The middle split marks the WER after self-training using 100 hours of Levantine. Finally, the bottom of each bar is the lower bound performance had the unlabeled data been transcribed. WER Recovery (the top section) is then the fraction of the entire bar covered by white. Column 3 has no gain and column 4 was negative.

7. Analysis and Conclusions

In Section 4, we noted the equivalent strengths between one hour of Levantine and ten hours of MSA as well as ten hours of Levantine and 1400 hours of MSA. However, the experiments in Section 6 showed that for pairs with similar initial WER, WER recovery was lower for the MSA models.

To understand why, we measured the bias of the acoustic models. For each AM, we produced a phonetic confusion matrix of the Levantine test set between the reference and hypothesis. Since these systems were graphemic, this was simply a character-level alignment. We then computed 1) the KL divergence of the unigram distribution of the hypothesis phonemes to the reference phonemes and 2) the mutual information (MI) between the hypothesis and reference phone given a recognition error (off the diagonal).

Each cell in the matrix is the joint probability of recognizing phone A when the reference was B. We also have the marginal probabilities of the reference, letting us compute the conditional probability of hypothesis A given reference B. We compute MI off the diagonal to control for model accuracy. An unbiased model should 1) have low divergence to the reference unigram distribution and 2) have low mutual information, since errors would be equally likely.

We make the following conclusions in Table 5. First, MI correlates very well with overall WER Recovery, with a coefficient of -0.83 across all eight conditions. As the bias increases, the effectiveness of self-training decreases. Compare columns 4 and 5. KL divergence does not correlate well, with a coefficient of -0.23. However, for a given pair of models with similar initial WER, lower KL divergence implies higher WER Recovery. Compare the initial WER and resulting Recovery for each pair of rows. Finally, initial WER is a poor predictor of WER Recovery, with a correlation coefficient of 0.26.

Since self-training effectiveness depends much more strongly on model bias than model accuracy, future work could consider techniques to trade reductions in model bias for an increase in error rate. Such techniques could potentially reduce the extreme mismatch between MSA and Arabic dialects. Despite both being labeled Arabic, they are different languages and it is clear that while acoustics need not be in-domain, a strong LM is still vitally important.

AM	LM	WER	Recov.	MI	KL
1 Lev	*Lev*	75.2	**53%**	.277	.058
10 MSA	*Lev*	75.3	**42%**	.281	.107
1 Lev	*MSA*	84.8	**25%**	.274	.074
10 MSA	*MSA*	82.5	**15%**	.297	.136
10 Lev	*Lev*	60.9	**17%**	.361	.028
1400 MSA	*Lev*	61.4	**16%**	.365	.065
10 Lev	*MSA*	70.1	**7%**	.345	.036
1400 MSA	*MSA*	69.8	**-18%**	.383	.118

Table 5 – *Impact of Bias on Self-Training.* Model bias explains why four pairs of models (contrast each pair of rows) with similar initial WER (col 3) had very different WER Recovery (col 4). Mutual information between hypothesis and reference phones when there is an error (col 5) negatively correlates with recovery. Higher KL divergence of the ASR unigram phoneme statistics to the reference (col 6) correctly predicts which model will have lower recovery.

8. References

[1] Loof, J., Gollan, C., Ney, H., "Cross language Bootstrapping for Unsupervised Acoustic Model Training: Rapid Development of a Polish Speech Recognition System", Interspeech 2009.

[2] Ma, J., Matsoukas, S., "Unsupervised Training on a Large Amount of Arabic Broadcast News Data", ICASSP 2007.

[3] Novotney, S. and Schwartz, R. "Analysis of Low-Resource Acoustic Model Self-Training", Interspeech, 2009

[4] Elmahdy, M., Gruhn, R., Minker, W., Minker, W., Abdennadher, Slim. "Cross-Lingual Acoustic modeling for Dialectal Arabic Speech Recognition", Interspeech 2010.

[5] Zhou, Q., Zitouni, I., "Arabic Dialectical Speech Recognition in Mobile Communication Services", Speech Recognition, 2008.

[6] Abdou, S., Arvizo, R., Atrash, A., Colthurst, T., Kao,CL, Kimball, O., Ma, J., Makhoul, J., Matsoukas, S., Prasad, R., Xu, D., Zhang, B. 2004. "The 2004 BBN Levantine Arabic and Mandarin CTS Transcription Systems" RT-04 Workshop, 2004

[7] Wang, L., Gales, M., Woodland, P., "Unsupervised Training for Mandarin Broadcast News and Conversation Transcription", ICASSP 2007.

[8] Gollan, C., Han, S., Schluter, R., Ney, H., "An Improved Method for Unsupervised Training of LVCSR Systems", Interspeech 2007.

[9] Nizar, H., "On Arabic and its Dialects", Multilingual Magazine. #81, Volume 17, Issue 5, 2006.

[10] Vergyri, D., Kirchhoff, K., Duh, K., Stolcke, A., "Morphology-Based Language Modeling for Arabic Speech Recognition", ICLSP 2004