Advances in Dialectal Arabic Speech Recognition: A Study Using Twitter to Improve Egyptian ASR

Ahmed Ali, Hamdy Mubarak, Stephan Vogel

Qatar Computing Research Institute
Qatar Foundation, Doha, Qatar
{amali,hmubarak,svogel}@qf.org.qa.org

Abstract

This paper reports results in building an Egyptian Arabic speech recognition system as an example for under-resourced languages. We investigated different approaches to build the system using 10 hours for training the acoustic model, and results for both grapheme system and phoneme system using Morphological Analysis and Disambiguation for Arabic (MADA). The phoneme-based system shows better results than the grapheme-based system. In this paper, we explore the use of tweets written in dialectal Arabic. Using 880K Egyptian tweets reduced the Out Of Vocabulary (OOV) rate from 15.1% to 3.2% and the Word Error Rate (WER) from 59.6% to 44.7%, a relative gain 25% in WER.

1. Introduction

Arabic Automatic Speech Recognition (ASR) is a challenging task because of the lexical variety and data sparseness of the language. Arabic can be considered as one of the most morphologically complex languages [1]. With more than 300 million people speaking Arabic as a mother tongue it is the 5th most widely spoken language. Modern Standard Arabic (MSA) is the official language amongst Arabic native speakers, in fact MSA is used in formal events, such as newspaper, formal speech, and broadcast news. However, MSA is very rarely used in day-to-day communication. Nearly all the Arabic speakers use Dialectal Arabic (DA) in everyday communication [2]. DA has many differences from MSA in morphology, phonology and lexicon [3]. A significant challenge in dialectal speech recognition is diglossia, in which the written language differs considerably from the spoken vernaculars [4]. The variance among different Arabic dialects such as Egyptian, Levantine or Gulf has to be considered similar to the variance among Romance languages [5]. There are many varieties of dialectal Arabic distributed over the 22 countries in the Arabic world, often several variants of the Arabic language within the same country. There is also the difference between Bedouin and Sedentary speech, which runs across all Arabic countries. However, in natural language processing, researchers have aggregated dialectal Arabic into five regional language groups: Egyptian, Maghrebi, Gulf (Arabian Peninsula), Iraqi, and Levantine [2][6].

A recent study [7] demonstrated that the use of the online User Generated Content (UGC) can help to improve the speech recognition by an average of 12.5% for the broadcast domain in French. This result on a high-resourced language like French motivates us to consider a similar approach for Egyptian dialectal Arabic, which has to be considered a low-resource language. In this paper, we report results for Egyptian Speech Recognition using limited speech data of 10 hours for training and 1.25 hours for development and testing. There has been recent interest in Egyptian speech recognition by [8][9][10]. This paper however differs from previous work by:

1. Investigating the best practices for writing Egyptian orthography, conducting experiments on both Acoustic Model (AM) and Language Model (LM), and releasing augmented Conventional Orthography for Dialectal Arabic [11] CODA guidelines for transcribing Egyptian speech.
2. Improving the dialectal Arabic speech recognition, and showing significant reduction in the word error rate using micro blog data, particularly tweets.
3. Comparing the dialectal tweet collection and the approach being used in classifying the tweets per country.

In addition, we release a tri-gram Egyptian language model, as well Egyptian lexicon that has less than 4% OOV on the test set.

2. Dialectal Arabic

Dialectal Arabic (DA) refers to the spoken language used for daily communication in Arab countries. There are considerable geographical distinctions between DAs within countries, across country borders, and even between cities and villages as shown in Figure 1[1].

Recent research [12][2][13] is based on a coarser classification of Arabic dialects into five groups namely: Egyptian (EGY), Gulf (GLF), Maghrebi (MGR), Levantine (LEV), and Iraqi (IRQ). Other dialects are classified as OTHER (see Figure 2). Zaidan [20] mentioned that this is one possible breakdown but it is relatively coarse and can be further divided into more dialect groups, especially in large regions such as the Maghreb.

[1] http://en.wikipedia.org/wiki/Arabic_dialects

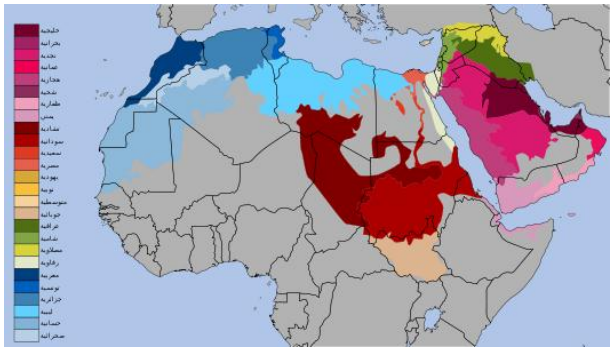

Figure 1: *Different Arabic Dialects in the Arab World.*

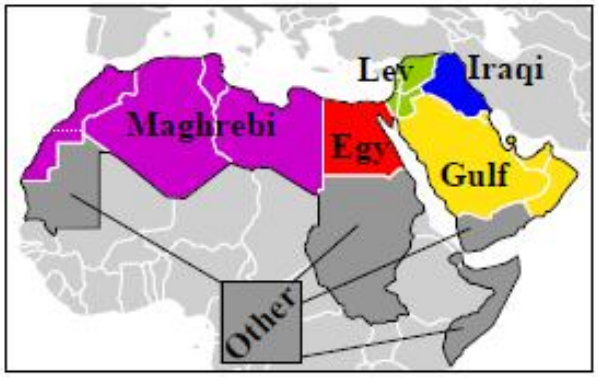

Figure 2: *Major Arabic Dialect Groups.*

3. Speech Data

3.1. Data Collection

The speech data used for this paper has been collected from Aljazeera Arabic channels, using two setups: satellite recording and internet video streaming from the Aljazeera.net website. The speech is recorded using 16 khz sampling rate. We looked at signals from both the satellite feed and online streaming, and the difference in quality is rather small and does not change anything in the quality of the audio as far as speech recognition is concerned.

A database of 200 hours has been collected over a period of six months in 2013 using the aforementioned setup. This data has been manually segmented to avoid speaker overlap, and avoid any non-speech parts such as music and background noise. These segments have a wide range of durations, from 3 seconds to 180 seconds. Speech segment were then classified as either Egyptian, Levantine, Maghrebi, Gulf, or MSA.

For the experiments described in this paper we used 12.5 hours of speech data classified to be in the Egyptian dialect, which was split into three subsets; training 10 hours, test set and development set 1.25 hours each. More details about the data are provided in Table 1.
We report the WER in this paper for both test set and development set; the first number is always for the test set and second number for the development set.

Table 1: *Speech Training Data Details.*

Duration	train(10h)	test(1.25h)	dev(1.25h)
#sentences	1385	147	176
#words	80K	9700	9809

3.2. Speech Transcription

As DA has no standard orthography or generally accepted writing convention, we investigated two approaches for manually transcribing Egyptian Speech data:
1) Verbatim transcription: The transcription is a faithful rendering of the speech without paying attention to language rules. E.g. the person name شفيق, $fyq[2] is typically pronounced by Egyptian native speakers as شفيء $fy, replacing the plosive /k/ in this context with a glottal stop hamza /A/. In this writing convention, the word will then be written as it has been pronounced, i.e. as $fy.
2) CODA-S (Augmented Coda for Speech Transcription): This transcription follows the CODA transcription guidelines [11], however, with some enhancements described below to address the needs for transcribing speech. In this case the transcription follows the language rules rather than the variant pronunciation.

CODA is mainly a framework for writing dialectal Arabic, but when working with transcribers it became apparent that some details were underspecified. We therefore augmented the CODA guidelines to make the rules clearer to the transcribers. We share these modified transcription guidelines and make them available on QCRI ALT web portal [3].

Here are some of the added explanations to the transcription guidelines. The shared document summarizes all cases by describing the case and providing samples of different writings in addition to the correct writing, as shown in Table 2, which shows one of the cases: Prefixes for future tense ("ح H" and "ه h") that are attached to present verbs, should be kept as they are without splitting from verbs

Table 2: *Examples of augmented CODA Guidelines.*

Various Writings	Correct Writing
حيبقي Hybqy, حيبقَا HybqA	حيبقَى HybqY
هَا يبقَى hA ybqY, هيبقي hybqy	هيبقَى hybqY

More rules have been added to cover cases not mentioned in the original CODA framework:

[2]Buckwalter encoding is used throughout the paper.
[3]http://alt.qcri.org/resources/speech/Egyptian/EgyptianTranscription_CODA.pdf

Split letter "ع E" that represents the preposition "علَى ElY" when concatenated to a noun. Ex: عَالأَرض EAl>rD → ع الأَرض E Al>rD.

Correct the suffix "و w", which is written instead of suffix "ه h". Ex: منو mnw → منه mnh, and عندو Endw → عنده Endh.

Restore "أ >" at the beginning of a present verb when the verb is prefixed by "ب b". Ex: بهزر bhzr → بأَهزر b>hzr.

Replace suffix "يَا yA" which indicates possession for the first person with suffix "يّ y ". Ex: فيَا fyA → فيّ fy .

Split negation article "مَا mA" in all cases. Ex: مفيش mfy$ → مَا فيش mA fy$, مَاعملتهَاش mAEmlthA$ → مَا عملتهَاش mA EmlthA$.

The guidelines also contain a new rule for punctuation marks and tags for hesitations or incomplete words, which is very important in speech transcription task. Ex: طيب ازَاي احلَى حَاجة يقولّك يد خفية Tyb Azay AHlY HAjp yqwllk yd xfyp → طيب ازَاي؟ أَحلَى حَاجة يقولّك يد خفية! Tyb <zAy? >HlY HAjp yqwl lk yd xfyp!
Finally, we added a long list of common words with different writings and the correct writing for each word. Ex: كدة kdp, كدَا kdA → كده kdh, and برضة brDp, بردو brdw → برضه brDh.

4. Dialectal Tweet Corpus

According to Twitter, the estimated number of Arabic microblogs is in excess of 15 million per day (private communication). To build a dialectal tweet corpus a multi-step procedure was used: 1) Arabic tweets were extracted by issuing the query lang:ar against the Twitter API[4].
2) Each tweet was classified as dialectal or not dialectal.
3) Dialectal tweets were mapped, if possible, to a country. If such a mapping was possible, the tweet was classified as being written in the dialect associated with that country according to Figure 2.

In more detail: To perform step 2, dialectal words were extracted from the Arabic Online Commentary Dataset (AOCD) described in [20]. Examples of words used in dialects: دي dy, عشَان E$An, هيك hyk, ايش Ay$, اكو Ako, شنو $nw, وَاش wA$ etc. As shown in [14], many of these dialectal words are used in more than one dialect. I.e. these words do not map a tweet uniquely to a dialect. For example the word "دي dy" is used in Egypt and Sudan, and the word "عشَان E$An" is used in Egypt and Arab Gulf countries etc.
If a tweet has at least one dialectal word, it was considered as dialectal tweet.

In step 3 user location in his/her profile was harvested and an attempt was made to identify the country with the aid of the GeoNames[5] geographical database. For examples: dialectal tweets with user locations like الريَاض AlryAD, Riyadh, KSA, الحجَاز AlHjAAz are mapped to Saudi Arabia and thereby to Gulf Arabic.

Applying the 3 filtering steps a corpus of size 6.5M tweets was collected during March 2014. The classification resulted in the following distribution: 3.99M tweets for Saudi Arabia (SA) (or 61% of the corpus size), 880K tweets for Egypt (EG) (13%), 707K tweets for Kuwait (KW) (11%), 302K for Arab Emirates (AE) (5%), etc. Tweets distribution is shown in Figure 3.

Using CrowdFlower[6] and 3 judges from Egypt we evaluated the accuracy for the automatic classification. Using 6,000 tweets classified as Egyptian, the achieved precision was 94%.

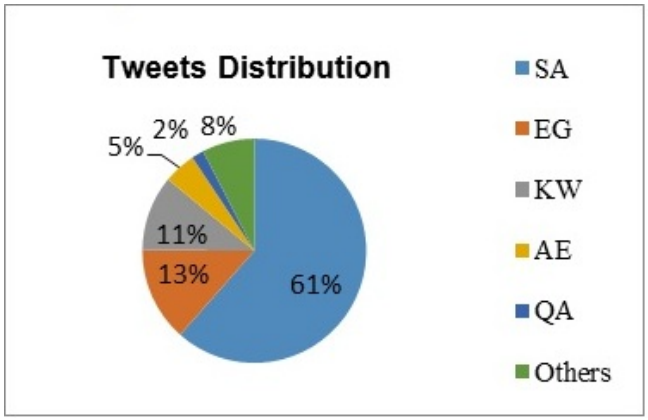

Figure 3: *Dialectal Tweets Distribution Percentages.*

5. Speech Recognition

This section describes the details of the speech recognition system, esp. the acoustic model training and the language models used in the experiments.

5.1. Language Modeling

Following [7] we wanted to test the impact of using tweets when building the language model for the speech recognition system. This leads to a number of questions: Is it better to use all dialectal tweets across the different dialects or is it better to use only the tweets in the matching dialect? How much do we gain by using more data? Does normalizing the tweets matter?

[4] http://dev.twitter.com/

[5] http://www.geonames.org/

[6] https://crowdflower.com

5.1.1. *Training Language Models*

We build standard trigram LMs with Kneser-Ney smoothing using SRI LM toolkit [18]. For interpolating LMs, the development set was used to tune the weight for the linear interpolation. In such cases we report only test set results, whereas in other cases we report numbers for both development and test set.

5.1.2. *Type/Token Ratios*

To answer the questions raised above we analyzed and compared several copora:

1) Speech data in verbatim format.
2) Speech data in CODA-S format.
3) Egyptian tweets without normalization.
4) Egyptian tweets with normalization, where we use the normalization method described in [19].
5) MSA sample, collected from the last 5 years of Aljazeera website.

One concern in statistical modeling is always data sparseness. When building language models data sparseness can be expressed in terms of type/token ratio. The higher the type/token ratio, the sparser the data becomes for LM training.

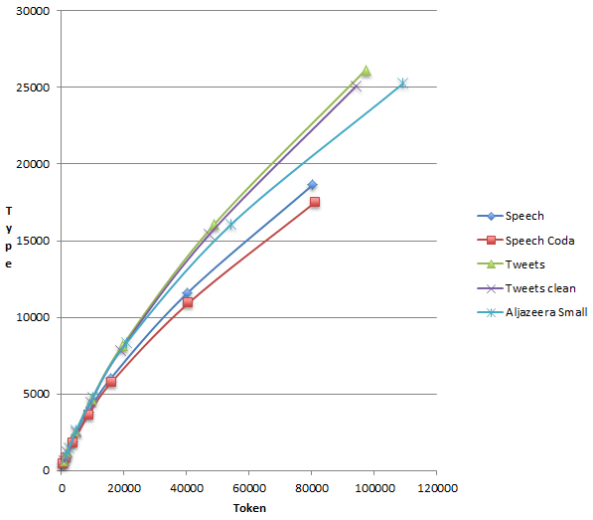

Figure 4: *Type Token Ratios for Various Text Samples.*

Figure 4 compares the type/token ratios across all the aforementioned corpora. This diagram shows how the vocabulary (number of types) grows as the corpus (number of tokens) grows. A number of observations can be made from this graph:

1) As expected, speech data shows a slower vocabulary growth compared to text data.
2) Using the CODA-S transcriptions reduces the type/token ratio, which should be benefitial for the performance of the speech recognition system.
3) The tweet corpus shows a higher type/token ratio than both speech and web-text corpora. This was not necessarily expected and could indicate that variants in writing are a major factor in dialectal tweets.
4) Normalizing tweets had only a minimal effect in improving the type/token ratio. Perhaps this could be improved with a tweet-optimized normalizer rather than the simple one [19] used here.

5.1.3. *Out of Vocabulary Rates and Perplexities*

In the next step we investigated the benefit of going towards larger vocabularies, also comparing Egyptian-only tweets (TweetsEGY) versus all dialectal tweets (TweetsALL). In this comparison we looked at OOV rates and at LM perplexities, which are based on interpolated LMs: one LM build on the speech corpus in CODA-S format, one LM build on a subset of the tweets.

As shown in Table 3 the Egyptian tweets have better results on the Egyptian test set. While the gains are not very big the difference actually grows with larger vocabulary sizes. For example the drop in OOV from TweetsAll to TweetsEGY is 15% for the 30K corpus, yet 20% for the 400k corpus. The perplexity drop is even more pronounced, going from 1.3% on the 30k corpus to 5.1% on the 400k corpus.

Table 3: *Compare tweetsEGY to tweetsAll LM.*

Data	Vocab	Perplexity	OOV
ALL	30K	1096	11.6%
EGY		1082	10.2%
ALL	50K	1269	9.4%
EGY		1242	8.4%
ALL	100K	1549	7.2%
EGY		1547	6%
ALL	200K	1891	5.3%
EGY		1834	4.2%
ALL	400K	2157	4.0%
EGY		2047	3.2%

Numbers reported in Table 3 are for the test data only, as we used it to tune the LM interpolation for the training data LM and tweet data LM.

The 400K interpolated LM and the corresponding lexicon have been released on QCRI ALT web portal [7].

5.2. Acoustic Modeling

Our acoustic models are trained with the standard 13-dimensional Cepstral Mean-Variance Normalized (CMVN)

[7] http://alt.qcri.org/resources/speech/Egyptian/egyVoc400K

Mel-Frequency Cesptral Coefficients (MFCC) features without energy, and its first and second derivatives. For each frame, we also include its neighboring +/-4 frames and apply Linear Discriminative Analysis (LDA) transformation to project the concatenated frames to 40 dimensions, followed by Maximum Likelihood Linear Transform (MLLT). We use this setting of feature extraction for all models trained in our system. Speaker adaptation is also applied with feature-space Maximum Likelihood Linear Regression (fMLLR).

Our system includes all conventional models supported by KALDI [15]: diagonal Gaussian Mixture Models (GMM), subspace GMM (SGMM) and Deep Neural Network (DNN) models. Training techniques including discriminative training such as boosted Maximum Mutual Information (bMMI), Minimum Phone Error (MPE), and Sequential Training for DNN are also employed to obtain the best number.

These models are all standard 3-states context-dependent triphone models. The GMM-HMM model has about 9K Gaussians for 1.8K states; the SGMM-HMM model has 4.5K states and 40K total substates.

We studied two ways of modeling the speech:
1) grapheme-based modeling, where each character represents a model. In this system we have 36 speech models plus one model for silence. The 36 models represent the 36 unique characters, which appear in our speech training data.
2) We also studied a phoneme-based system, where we preprocessed the training text using the Morphological Analysis and Disambiguation for Arabic (MADA) toolkit [16], which has been used to build a vowelized dictionary. A rule-based Vowelized to Phonetized (V2P) mapping was then used to generate the final lexicon. The phoneme system has 36 phones: 35 speech phonemes and one phoneme for silence. More details are provided in [21].

It is worth mentioning that MADA was developed for MSA and therefore may not the best tool for pre-processing dialectal Arabic. We learnt about MADAMIRA, which merges MADA [16] and AMIRA [17]. This tool provides linguistic information such as tokenization, diacritization, and part-of-speech tagging for each Arabic word received in corpus, which supports Egyptian text. However due to license restrictions, we were unable to use it in our experiments.

Table 4: *Comparing grapheme-based and phoneme-based systems, both with CODA-S transcriptions.*

Train data LM	Grapheme	Phoneme
1st pass WER	62.47%	51.27%
	68.41%	58.14%
2nd pass WER	59.63%	47.73%
	64.68%	53.73%

Table 4 shows that the phoneme system outperforms the grapheme system substantially with 20% relative reduction in WER. One reason behind this gain is that in Arabic the correspondence between phoneme and grapheme is weak due to the short vowels, which are not written. Consequently, mapping each grapheme as a unit will fall short to model in the GMM the different variants occuring in the training data. Also, the grapheme system needs more contexts to disambiguate between phonemes.

Although this is a nice reduction in WER, the range of the error is still high, which is not a surprise given the high OOV rate and perplexity, which raises the question: is it possible to use the Egyptian tweets to build better language model to improve the dialect speech recognition? This will be addresssed in the experiments described in the next section.

6. Experiments

6.1. CODA-S and Verbatim Comparison

In an attempt to depict which approach is more appropriate to use for transcribing the Egyptian speech, we used two techniques to evaluate best approach by reporting OOV, Perplexity (PP) and ASR system and report WER.
a- Evaluating using Language Model only (LM) the test and dev set with the collected Egyptian tweets, and report OOV and PP. We used Egyptian tweets to build trigram LM, more details about Egyptian tweets in section 4, and LM in section 5. In Table 5, we report PP and OOV for both CODA-S and verbatim transcription convention. The first value refers to test set and the second to dev set. CODA-S is getting better results in both PP and OOV.

Table 5: *PP & OOV for CODA-S and Verbatim. (Type: 395K words, Tokens: 9.5M words)*

	Verbatim	CODA-S
PP	6729	5837
	6978	6031
OOV	6.8%	4.7%
	6.3%	4.6%

b- Building Grapheme based speech recognition, and report WER.
For the speech, we investigate WER at different stages of the Acoustics Model (AM) process, however, we report only the WER at the very last stage which is Deep Neural Network (DNN) with Minimum Phoneme Error (MPE). We report the WER at first pass and the second pass, again the first value refers to test set and the second to dev set. More details about the speech recognition system are covered in section 5.

Table 6 shows that the number of words in the verbatim transcription is 80.4K words, while the total number of words in the CODA-S transcription is 81K words. Although there is a small increase in the amount of words, there is a decrease in the vocabulary size from 18.6K words in verbatim

text to 17.5K in the CODA-S text, which represents nearly 6% reduction. This is due to more consistency in writing the text which consequently reduces the sparseness in the text. It is worth mentioning that the WER comparison may not be fair measure by itself as it is impacted by Acoustic Modeling (AM) as well as Language Modeling (LM). Having said that, in this setup we used grapheme based AM approach in both systems to be consistent with AM, and reduce the acoustic influence on the conclusion. Also, best WER does not necessary mean the best orthographic representations. But, authors found WER could be an extra measure to consider. It is clear from Table 6 that WER, PP, and OOV in the CODA-S format are consistently outperforming the verbatim transcription, which was a go-ahead signal for us to use the CODA-S as baseline for all our further experiment.

Table 6: *WER for CODA-S and verbatim*

	Verbatim	CODA-S
Token	80.4K	81K
Type	18.6K	17.5K
1st pass WER	64.78%	62.47%
	69.77%	68.41%
2nd pass WER	61.20%	59.63%
	65.98%	64.68%
Perplexity	957	862
	976	855
OOV	16.7%	15.1%
	16.6%	15.4%

6.2. Grapheme versus Phoneme based System

At this stage, we see the best AM system is the phoneme system and the best LM is the 400K vocabulary for the interpolated LM. So, the next step is to use an interpolated LM with the phoneme system and expect that the gain in both LM and AM will propagate to the final system. One challenge in doing so is that we have to pre-process tweetsEGY by MADA to generate a lexicon for the phoneme system. However, as already mentioned, MADA is not the best tool to vowelize Egyptian dialectal Arabic. So, we compared both systems in Table 4 with the 400K interpolated LM and get the WER as shown in Table 7.

Table 7: *Compare Grapheme and Phoneme CODA-S Systems Using 400K Interpolated LM.*

Tweet interpolated LM	Grapheme	Phoneme
1st pass WER	47.31%	56.22%
	54.26%	62%
2nd pass WER	44.71%	52.73%
	50.62%	58.60%

We see the LM helped the Grapheme system substantially and reduced the WER by more than 25% relative in test set (from 59.63% shown in Table 4 to 44.71% shown in Table 7), and more than 21% relative reduction in development set (from 64.68% shown in Table 4 to 50.62% shown in Table 7) in the development set. In the phoneme system this gain from the tweets LM has not only vanished, but we get an increase in error rate by 10% and 9% relative in test set and development set. At this time we assume that this increase in WER stems from the fact that we do not have access to a reasonable Egyptian vowelizer, nor a nice tool that can convert dialectal Egyptian tweets into the CODA-S format.

6.3. TweetsEGY versus TweetsAll

We have also investigated the importance of doing dialect detection for the tweets, and compared the WER using the Egyptian tweets versus random selection for any Arabic tweets. We see in Table 8, the dialect identification does give us some mileage. We can see a difference in WER of about 3 points absolute across both decoder passes. We report the WER on the test set only as the development set has been used to tune lambda for the linear interpolation.

Table 8: *Compare tweetsEGY WER versus tweetsAll.*

Interpolated LM	Grapheme EG	Grapheme All
1st pass WER	47.31%	50.3%
2nd pass WER	44.71%	47.2%

7. Conclusion

Dialectal Arabic speech recognition is a challenging task when analyzing the available resources. In this paper, we report significant reduction in WER by approaching different aspects of the challenge: we standardize augmented CODA guidelines for transcribing Egyptian speech to reduce the impact of diglossia. We used tweets for improved vocabulary coverage and significantly reduced WER. Using specifically tweets classified as being written in the Egyptian dialect gave lower WER than using tweets across all dialects. We released the language model as well as the lexicon used in this paper. In future work, we plan to work on better dialectal vowelizer to be able to generate lexicons for different dialects. We will also investigate how to convert tweets into the CODA-S format automatically. Given the benefit of being dialect specific, we will analyze tweets that are not mapped to countries, and study using tweet location in addition to user location to enhance mapping accuracy, also enrich the dialectal words list and assign each dialectal word to a country or a set of countries.

8. References

[1] Diehl, F. et al., "Morphological decomposition in Arabic ASR systems", Computer Speech & Language, 26(4), pp.229243, 2012.

[2] Cotterell, R. & Callison-Burch, C., "A Multi-Dialect, Multi-Genre Corpus of Informal Written Arabic", The 9th edition of the Language Resources and Evaluation Conference, Reykjavik, Iceland: European Language Resources Association, 2014.

[3] Habash, N., Eskander, R. & Hawwari, A., "A Morphological Analyzer for Egyptian Arabic", pp.19, 2012.

[4] Elmahdy, M., Hasegawa-Johnson, M. & Mustafawi, E., "Development of a TV Broadcast Speech Recognition System for Qatari Arabic", The 9th edition of the Language Resources and Evaluation Conference (LREC), Reykjavik, Iceland, 2014.

[5] Holes, C., "Modern Arabic: Structures, Functions, and Varieties", 2004.

[6] Al-Sabbagh, R. & Girju, R., "YADAC: Yet another Dialectal Arabic Corpus", LREC, pp. 28822889, 2012.

[7] Schlippe, T. et al., "Unsupervised language model adaptation for automatic speech recognition of broadcast news using web 2.0", F. Bimbot et al., eds. INTERSPEECH, ISCA, pp. 26982702, 2013.

[8] Biadsy, F., Moreno, P.J. & Jansche, M., "Googles Cross-Dialect Arabic Voice Search", IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2012), pp. 44414444, 2012.

[9] Al-Shareef, S. & Hain, T., "CRF-based Diacritisation of Colloquial Arabic for Automatic Speech Recognition", INTERSPEECH, 2012.

[10] Mousa, A.E. et al., "Morpheme-Based Feature-Rich Language Models Using Deep Neural Networks for LVCSR of Egyptian Arabic Human Language Technology and Pattern Recognition", Computer Science Department IBM T . J . Watson Research Center, Yorktown Heights, NY 10598, pp.84358439, 2013.

[11] Habash, N., Diab, M.T. & Rambow, O., "Conventional Orthography for Dialectal Arabic", LREC, pp. 711718, 2012.

[12] Zbib, R. et al., "Machine Translation of Arabic Dialects", In Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 12, Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 4959, 2012.

[13] Bahari, M.H. et al., "Non-negative factor analysis for GMM weight adaptation", IEEE Transactions on Audio Speech and Language Processing, 2014.

[14] Mubarak, H. and Darwish K., "Using Twitter to Collect a Multi-Dialectal Corpus of Arabic", Arabic NLP Workshop, EMNLP-2014, 2014.

[15] Povey, D. et al., "The Kaldi Speech Recognition Toolkit", IEEE Signal Processing Society, 2011.

[16] Habash, N., Rambow, O. & Roth, R., "MADA+TOKAN: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS tagging, stemming and lemmatization", In Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR), Cairo, Egypt. pp. 102109, 2009.

[17] Diab, M., "Second generation AMIRA tools for Arabic processing: Fast and robust tokenization, POS tagging, and base phrase chunking", 2nd International Conference on Arabic Language Resources and Tools, 2009.

[18] Stolcke, A. & others, "SRILM-an extensible language modeling toolkit", INTERSPEECH, 2002.

[19] Darwish, K., Magdy, W. & Mourad, A., "Language processing for Arabic microblog retrieval", In Proceedings of the 21st ACM international conference on Information and knowledge management. pp. 24272430, 2012.

[20] Zaidan, O.F. & Callison-Burch, C., "Arabic Dialect Identification", Computational Linguistics, 40(1), pp.171202, 2014.

[21] Ali, A. et al., "A Complete KALDI Recipe for Building Arabic Speech Recognition Systems", Spoken Language Technology Workshop (SLT), IEEE, 2014.