Multi-Reference Evaluation for Dialectal Speech Recognition System: A Study for Egyptian ASR

Multi-Reference Evaluation for Dialectal Speech Recognition System: A Study for Egyptian ASR

Ahmed Ali[1,2], **Walid Magdy**[1], **Steve Renals**[2]
[1]Qatar Computing Research Institute, HBKU, Doha, Qatar
[2]University of Edinburgh, Edinburgh EH8 9AB, UK
{amali,wmagdy}@qf.org.qa, {ahmed.ali, s.renals}@ed.ac.uk

Abstract

Dialectal Arabic has no standard orthographic representation. This creates a challenge when evaluating an Automatic Speech Recognition (ASR) system for dialect. Since the reference transcription text can vary widely from one user to another, we propose an innovative approach for evaluating dialectal speech recognition using Multi-References. For each recognized speech segments, we ask five different users to transcribe the speech. We combine the alignment for the multiple references, and use the combined alignment to report a modified version of Word Error Rate (WER). This approach is in favor of accepting a recognized word if any of the references typed it in the same form. Our method proved to be more effective in capturing many correctly recognized words that have multiple acceptable spellings. The initial WER according to each of the five references individually ranged between 76.4% to 80.9%. When considering all references combined, the Multi-References MR-WER was found to be 53%.

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 one of the most morphologically complex languages (Diehl et al., 2012). With more than 300 million people speaking Arabic as a mother tongue, it is counted as the fifth 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 newspapers, formal speech, and broadcast news.

Nevertheless, MSA is rarely used in day-to-day communication. The vast majority of Arabic speakers use Dialectal Arabic (DA) in everyday communication (Cotterell and Callison-Burch, 2014). DA has many differences from MSA in morphology, phonology and the lexicon. A significant challenge in dialectal speech recognition is diglossia, in which the written language differs considerably from the spoken vernaculars (Elmahdy et al., 2014). Variance among different Arabic dialects such as Egyptian, Levantine or Gulf has been considered similar to the variance among Romance languages (Holes, 2004). There are many varieties of dialectal Arabic distributed over the 22 Arabic countries, often several variants of the Arabic language within the same country.

In natural language processing (NLP), researchers have aggregated dialectal Arabic into four regional language groups: Egyptian, Maghrebi, Gulf (Arabian Peninsula), and Levantine (Cotterell and Callison-Burch, 2014; Al-Sabbagh and Girju, 2012; Darwish and Magdy, 2014).

Most ASR systems are trained and tuned by minimizing WER, which counts word errors at the surface level. It does not consider the contextual and syntactic roles of a word, which are often critical for tasks like Machine Translation (MT), particularly in the end-to-end Speech Translation (ST) scenarios.

In a study by (He et al., 2011) , they showed that WER is not the optimal metric for a speech recognizer trained for a speech translation task. They developed a BLEU-optimized approach for training the scale parameters of a log-linear based speech translation system. In their study, they got better results using the new measure, although WER were found to be higher in the intermediate step of the speech recognition.

Dialectal Arabic can be viewed as an example of a language with no orthographic rules, since

Proceedings of the Second Workshop on Arabic Natural Language Processing, pages 118–126,
Beijing, China, July 26-31, 2015.

there is no academies in DA nor enough amount of language resources, such as no standard lexicon or clear rules for writing. In a study by (Habash et al., 2012) in which they presented Conventional Orthography for Dialectal Arabic CODA, they explain the design principles of CODA and provide description of CODA, and use the Egyptian dialect as an example, which has been presented mainly for the purpose of developing DA computational models.

In a similar study by (Ali et al., 2014a), they studied the best practices for writing Egyptian orthography. They conducted experiments on both Acoustic Model (AM), Language Model (LM), and guidelines for transcribing Egyptian speech. They released guidelines for transcribing Egyptian speech for what is called augmented Conventional Orthography for Dialectal Arabic augmented-CODA. They also reported gain in Egyptian speech recognition when augmented-CODA is followed in transcribing Egyptian speech data.

Unlike previous work by (Habash et al., 2012; Ali et al., 2014a), where they studied the best practices for writing DA, in this paper, we propose an evaluation method that accepts the variations in transcribing dialectal Arabic. We use multiple references, up to five different transcriptions per utterance, to evaluate the performance of the speech recognition engine. The main idea is to learn from the crowd and use multi-references to vote for each word in the recognized output. This is, in a way, similar to BLUE score used in MT, where multiple translation could be accepted for one source sentence. Here, we submit our speech data on a crowdsourcing platform, and ask for five different transcriptions for each speech segment. These five transcriptions typically capture the different acceptable variations of the Arabic dialect, where we then use them as our multiple references to calculate *multi-reference* WER (MR-WER).

The rest of the paper is organized as follows: In section 2, we describe dialectal speech recognition; section 3, we discuss the details of the multi-reference WER, and the proposed method evaluate dialectal ASR; section 4, we elaborate the data used in this experiment; section 5, we discuss the experiment and the results; and section 6 is for conclusion and future work.

2 Dialectal Speech Recognition

Large Vocabulary Speech Recognition (LVCSR) has been studied thoroughly in well-developed languages such as English, French, and Spanish. Also, MSA has obtained good results over the past decade as a result of GALE project, as well as more attention is paid to Arabic Broadcast domain(Diehl et al., 2012; Mangu et al., 2011; Cardinal et al., 2014; Ali et al., 2014b).

Dialectal Arabic ASR could be seen as under-resourced as it is lacking the basic component to have a decent system, such as enough labeled speech data for training, a lexicon, and a Natural Language Processing (NLP) pipeline for phonetic systems. Moreover, DA Arabic lacks standard orthography for writing. The absence of clear definition for right and wrong spelling has led to many representations for each word.

In our Arabic ASR, we use a grapheme-based system using sequential Deep Neural Network for the acoustic modelling. Although, conventionally, a phoneme system always outperforms a grapheme system, so a valid question is why do we choose grapheme system here?

We have found that WER in the grapheme system has increased by less than 1% relative to conversational speech compared to the phoneme system, which could be explained as conversational speech being mainly dialectal Arabic in most cases, and grapheme models will outperform phoneme models. Mainly, the NLP pipeline for phonetic system is not mature enough for dialectal Arabic, and is still facing challenges such as diacratization, and phonetization. The other amusing feature in the grapheme system is having a 1:1 ratio between the number of types, and the number pronunciation in the lexicon, compared to 1:4 in the phoneme-based system. This enables us to increase the lexicon size from 500K words to more than 1.2M words for the same text in the Language Model (LM) with small impact on memory. This has reduced the Out Of Vocabulary (OOV) from 3.9% to 2.5%, which also enables us to have more coverage for dialectal words that have not been measured precisely at this stage.

3 Applying Multi-Reference Evaluation for ASR

In this section, we discuss the reason for proposing our new methodology of evaluating ASR, particularly DA ASR, using multiple references instead

of the standard method of using only one reference. In addition, we introduce our methodology for applying multi-reference to ASR evaluation.

3.1 The Concept of Multi-Reference Evaluation

One of the tasks that uses multi-reference evaluation is Machine Translation (MT). The main reason here is that many translations in the target language are fully valid for a given sentence in the source language. Thus, the MT research community found it more appropriate when evaluating an MT system to compare the automatic translation to more than one possible manual reference translations, typically translated by different language experts, to have a less biased evaluation to one translation. Therefore, most of the MT evaluation scores are designed to accept multiple references (Papineni et al., 2002).

ASR is treated differently, since the speech recognition is seen to have a single exact match to a specific string, and one reference should be sufficient to transcribe or judge what is spoken in the speech segment. This assumption is valid in most of the spoken languages. However, for languages with no standard orthographic representation such as Dialectal Arabic, there are many different ways to write a given spoken word. Table 1 shows an example for an Egyptian speech segment, which presents the transcription of one sound track from four different transcribers. As shown, many of the words presented has various spellings among the four transcribers. In addition, there are some words that are written by some transcribers but neglected by the others, such as the word "اه"(Ah) and "يعني"(yEny), that could be seen by some people as noise or filler and not worthy of writing. The variations in spelling the same words are clear in the shown example, such as {"ده"(dh),"دا"(dA)} and {"احنا"(AHnA),"نحن"(nHn), "إحنا"(AHnA)}.

Table 2 presents some additional samples of Arabic dialect words that have multiple acceptable spellings. These examples illustrate the problem of comparing an ASR output to only one reference that picks one of many possible spellings of a dialect Arabic word.

Accordingly, we propose introducing a multi-reference evaluation methodology for ASR tasks that targets languages with no standardized orthography. Similar to BLEU score in MT, multi-reference increases the likelihood of accepting an automatic translation (speech recognition), if any of the manual translations (transcriptions) agreed with it in some portions.

3.2 Multi-References Alignment to Recognized Speech Text

Our approach here is to extend the current alignment used when performing ASR evaluation between recognized text and one reference text to allow alignment between the recognized text and N references.

For a recognized text Rec = $\{w'_1, w'_2, ... w'_{|Rec|}\}$, and a set of N references: $Ref1$ = $\{w_{11}, w_{12}, ... w_{1|Ref1|}\}$ to $RefN$ = $\{w_{N1}, w_{N2}, ... w_{N|RefN|}\}$, we perform the following steps:

- For each word in Rec, list all the words in $Ref1$ to $RefN$ that are aligned to it. Note, that some references may not include any corresponding word for some of the words in Rec, which is counted as an insertion. The output of this process will be an array of size N of reference words for each recognized word.
  e.g.: $w'_3 \rightarrow [\ w_{12}, w_{23}, <\text{INS}>, ..., w_{N4}\ ]$
- The previous step effectively captures insertions, substitutions, and correct recognition. However, deletions would not be handled, since there is no corresponding word in the Rec to the deleted words in the reference. In addition, different number of deletions could exist across different references. To map deletions effectively across multiple references, for each reference, we map any non-aligned word to the recognized text to a "deletion pointer" (<DEL>) with a counter to the position of the last aligned word in Rec. For example, if two deletions are detected for one reference after 3 aligned words with Rec, the words in reference would be mapped to {"03-01 <DEL>", "03-02 <DEL>"} in the Rec. If another deletion is detected after the fifth word in Rec, it will be mapped to "05-01 <DEL>". For deletion pointers that are mapped to some of the references only, those reference that has nothing deleted would be assigned to "NULL". See Table 3 as an example.

For example shown in Table 1, the ASR system produced the following sentence:

Different Transcription
نعم اه طبيعي إن دة أصلاً إحنا في وضع غير قانوني بالمرة غير دستوري بالمرة وضع nEm Ah TbyEy <n dp >SlAF <HnA fy wDE gyr qAnwny bAlmrp gyr dstwry bAlmrp wDE
نعم اه طبيعي دا أصلا يعني أحنا فى وضع غير قانوني بالمرة غير دستوري بالمرة اه وضع nEm Ah TbyEy dA >SlA yEny >HnA fY wDE gyr qAnwny bAlmrp gyr dstwry bAlmrp Ah wDE
نعم نعم اه هو طبيعي ده اصلا احنا في وضع غير قانوني بالمره غير دستوري بالمره وضع Ah hw TbyEy dh ASlA AHnA fy wDE gyr qAnwny bAlmrh gyr dstwry bAlmrh wDE
نعم هو طبيعى دا أصلا يعنى نحن فى وضع غير قانونى بالمره غير دستورى بالمره وضع nEm hw TbyEY dA >SlA yEnY nHn fY wDE gyr qAnwnY bAlmrh gyr dstwrY bAlmrh wDE

Table 1: Different transcriptions for the same utterance

أعطى بإن دا أصلا يعني إحنا في وضع غير قانوني" بالمر غير دستوري بالمر واضح أه فيه انقلاب".

The alignment algorithm with the four references would produce the alignments shown in Table 3. As shown, now each word in the recognition is aligned to N references, which maximize the likelihood of finding a possible match that is accepted by one of the references.

3.3 Calculating MR-WER

Using the multi-aligned references, the number of correct, insertions, substitutions, and deletions are calculated as follows:

- **C** (Correct): is the number of recognized words that has a match in any of the aligned reference words.
- **S** (Substitutions): is the number of recognized words that has alignment to at least one reference words, but none of them matches it.
- **I** (Insertions): is the number of recognized words that is not aligned to any reference word. i.e. all corresponding alignments are "<INS>".
- **D** (Deletions): is the number of "<DEL>" instances in the Rec that has no "NULL" alignment in any of the references. The main reason for not counting deletions that has no corresponding word in one of the references is for the following assumptions: if one of the reference transcriptions decided that one of the spoken words is not worth transcribing, then the ASR should not be penalized for missing it. We can refer to example like the word "اه"(Ah) and "يعني"(yEny) where some of the transcribers considered them as a noise, and they decided not to write it.

Based on the counts of C, S, I, and D, MR-WET is calculated according to the following equation:

$$WER = \frac{S + D + I}{(S + D + C)}$$

In the case of multi transcriptions per reference, the length of the transcription varies from one reference to another which means that the deletion count is different among different transcriptions as shown in Table 3. By look at examples in this table, we can see that first reference has 16 words, the second one has 17 words, the third 17, and the fourth 16, we can see the number of words varies from one example to another. More specifically, the second transcriber decided to add the word "اه"(Ah) which none of the other three references considered it as a valid word. We can also see the third reference decided to add the word "نعم"(nEm) at the beginning which no one else added.

By applying the same WER equation mentioned above, we can see that reference 1 will have WER

Translation	Valid Spellings	Buckwalter
He was not	ماكانش	mAkAn$
	ماكنش	mAkn$
	ما كانش	mA kAn$
	مكنش	mkn$
I told him	قولتله	qwltlh
	قولت له	qwlt lh
	قلتله	qltlh
	قلت له	qlt lh
In the morning	على الصبح	ElY AlSbH
	علي الصبح	Ely AlSbH
	ع الصبح	E AlSbH
	عالصبح	EAlSbH
	عصّبح	ESbH

Table 2: Sample of phrases with multiple valid spellings

75%, reference 2 58%, reference 3 87%, and finally reference 4 will have 78% WER.

The MR-WER will have better results than any of the references distinctively, the MR-WER will be calculated as follow: :

$$MR - WER = \frac{6 + 1 + 2}{(6 + 1 + 10)}$$

Which is 52.6% WER, obviously, this is lower than the the lowest WER in any of the references.

4 Data

The data used in our experimentation comes from Broadcast News BCN domain; particularly, Aljazeera Arabic news channel. The nature of the data is debates and news programs which were uploaded to Al Jazeera in the duration between June 2014 and January 2015. All the speech data have gone through the pre-processing steps before being submitted to the used crowdsource platform [1] for transcription. Pre-processing included: removing non-speech audio such as music or white noise, followed by speaker segmentation and clustering, diarization, and speaker linking within the same episode. In addition to this, a dialect classification was performed using human computation, which also occurred via crowdFlower platform. Utterances underwent dialect classification by 3-9 annotators per audio file into five broad Arabic dialect groups: Modern Standard Arabic (MSA), Egyptian (EGY), Levantine (LEV), North African/Maghrebi (NOR), and Gulf (GLF). For the current study, we used audio segments which had been classified as EGY with at least 75% agreement between annotators.

[1] http://www.crowdFlower.com

Index	Rec	$Ref1$	$Ref2$	$Ref3$	$Ref4$
(00-1)	<DEL>	NULL	NULL	نعم	NULL
(00-2)	<DEL>	نعم	نعم	نعم	نعم
(01)	أعطى	اه	اه	اه	هو
(02)	بإن	طبيعي	طبيعي	هو	طبيعى
(03)	دا	إن	دا	طبيعي	دا
(04)	أصلا	دة	أصلا	ده	أصلا
(05)	يعني	أصلاً	يعني	اصلا	يعنى
(06)	إحنا	إحنا	أحنا	احنا	نحن
(07)	في	في	فى	في	في
(08)	وضع	وضع	وضع	وضع	وضع
(09)	غير	غير	غير	غير	غير
(10)	قانوني	قانوني	قانوني	قانوني	قانونى
(11)	بالمر	بالمرة	بالمرة	بالمره	بالمره
(12)	غير	غير	غير	غير	غير
(13)	دستوري	دستوري	دستوري	دستوري	<INS>
(14)	بالمر	<INS>	<INS>	<INS>	<INS>
(15)	واضح	<INS>	<INS>	<INS>	<INS>
(16)	أه	<INS>	بالمرة	<INS>	دستورى
(17)	فيه	بالمرة	اه	بالمره	بالمره
(18)	انقلاب	وضع	وضع	وضع	وضع
WER	MR:52%	75%	59%	88%	68%

Table 3: Alignment applied between a recognized text (Rec) and four different references

In this study, Egyptian data was chosen as a test case to take advantage of the fact that the classification pre-processing showed us that approximately 40% of users of the crowdsource platform in the Arab world are located in Egypt, meaning that focusing on EGY audio and Egyptian annotators allowed us to complete transcription fairly quickly. Furthermore, there were significantly more audio segments classified with high levels of inter-annotator agreement as EGY when compared to other dialect categories. Finally, EGY as a category contains a potentially less diverse set of dialects than a more geographically spread regional category.

We have asked for five references for 2765 speech segments (utterances), representing 4.8 hours, with speech segments of an average length between 4-6 seconds. Our results are based on these five files or sometimes mentioned as five references. This does not necessary mean five different annotators. It is mainly five transcription representations that have come from more than one annotator. Allocating transcription tasks to annotators and randomize the data to make sure one single editor did not write the same sentence more

than once was managed through the crowdFlower platform.

5 Experimentation

Our experiments are designed to address the following research questions:

1. How many references should be used in the multi reference
2. What is the inter-reference agreement? How good is the crowdsource data? Do we need to filter bad transcription for the MR-WER evaluation?
3. How many times do we need to see correct word to count it correct?

5.1 Number of References

We have evaluated the speech recognition using various number of N reference transcriptions, where N ranged from 1 to 5. We have used all the combinations between reference transcriptions in cases when $N > 1$ to validate our findings. As shown in Table 4, for every experiment, we report the minimum, maximum and average MR-WER for each number of transcriptions we use. We conclude from this experiments two findings:

1. The WER reduces considerably when we increase the number of transcriptions from one reference to five references, and may be there is potential to reduce the WER more if there are more transcriptions (although we can see the reduction in MR-WER between four and five references is not significant). The multi-reference evaluation has taken the error from an average WER of 80.8% to 53.5%. The 33% difference in performance are possibly happening due to various ways of writing DA not really due to bad ASR.
2. The variance in WER reduces noticeably when we increase the number of references. For example if you look at Figure3, the WER for five single references varies from 76.4% to 80.9%, with a absolute difference of 3.7% which is high error margin, for two references, the absolute difference is 3.4%, and in three references is 1.9% and in five references, it is only 0.5%. Obviously, we have only one WER for five transcriptions, as there is no combination between multiple transcriptions. Possible explanation for this nice reduction is error margin is that multi reference is capable to capture some of the variations in transcription, and make the reported error rate more robust.

# Re.f	One	Two	Three	Four	Five
Min.	78.5%	65.5%	59.3%	55.8%	53.5%
Av.	80.8%	66.7%	60.0%	56.0%	
Max.	82.2%	68.9%	61.2%	56.3%	
# Exp.	5	10	10	5	1

Table 4: MR-WER for various number of references per experiment.

5.2 What is the inter-reference agreement

The transcribed data is suffering from a very limited quality control that have been applied to it, which raised an important question: what is the inter-annotator agreement in this transcription task? This is a difficult question to ask in language with no clear orthographic rules. In most of the cases, if we consider exact string matches between different transcriptions even if it is perfect, the inter-annotator agreement is almost zero as shown in Table 4.

We evaluated WER for every transcription file with the other four files. For every utterance in each reference, there will be four WER for the same utterance in the other four files, the WER will be averaged. Each file has 2760 utterances, corresponding to 4.8 hours. We split the 2760 averaged WER values into four bins, WER 0-25%, WER 25-50%, WER 50-75% and anything more than 75%. We plot the results as seen in figure 1. It is clear from the aforementioned figure that there is a great deal of missmatch between the five references. Partially, this is due to bad transcription coming from some of the crowd source contributors, that we did not apply quality check at this stage.

As an attempt to quantify the bad transcriptions issue, and their impact on our experiment, we did some cleaning up for the data by removing any utterance that has more than 90% WER across the other four annotators. This is a very simple way assuming the majority of the transcription are correct, and may be invalid in a case where there is a single good transcription and the other four are bad, which has not been noticed in our corpus.

This experiment has reduced the number of utterances as shown in Table 5, so $Ref1$ has gone from 2765 utterances to 1824, $Ref2$ from 2765 to 2160 .. etc. Also, we plot the clean data as shown in 2 To evaluate the impact of data cleaning on the MR-WER, we run the same algorithm as explained in section 3.3 on the clean data, and we found that the MR-WER for the five references actually has increased from 53.5% to 54.2%.

This is an interesting finding to say that by cleaning the evaluation data, the MR-WER has not got any better, which could be explained because removing the potentially noisy data did not impact the MR-WER rather than removing some of the examples that could help in finding alternatives for Dialectal words. Also, it is fair to say that the proposed method is robust for the noisy data.

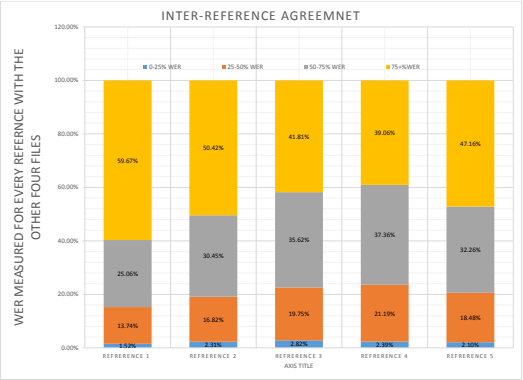

Figure 1: Inter-reference agreement for the full data

$Ref1$	$Ref2$	$Ref3$	$Ref4$	$Ref5$
1824	2160	2351	2414	2193

Table 5: Number of utterances per file after removing outlier transcriptions.

5.3 Counting Correct Words

In the case of single reference, the algorithm will loop over the solo reference, and check each word; insertion, deletion, substitution or correct. However, in the MR scenario, someone can argue that the algorithm in acting like cherry picking and looking for correct word in any of the references to make the WER look better rather than validating these findings. Basically, the spirit for this algorithm is try to find the recognized word in any

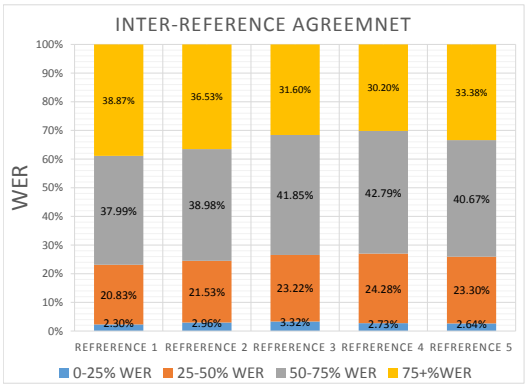

Figure 2: Inter-reference agreement after removing outlier transcriptions

of the references, obviously minding the position in text as explained in section 3.

To address this concern, we explore the impact in MR-WER when the algorithm asks for more than one evidence that a word is correct, i.e the same word occurred in same position in more than one reference. We evaluated correct word counting in 1+ (standard), 2+ and 3+ occurrences. Obviously, we apply N number of times seeing the word correct if there is N number of references or more.

We can see it clear in the alignment algorithm as shown in Table 3. The proposed MR-WER for the example in this table is 52.6%. In row index 03, the word "دا" (dA) will count correct for count 1+, and 2+, but not 3+. Row index 06, the word "إحنا" (ehnA) will count correct only in the 1+ count..etc.

The MR-WER for the example of at least two correct will be: 56.25% as the number of correct will reduce to 9 instead of 10. Same in the case of three correct examples or more, the MR-WER will be 64.28% as the number of correct examples will be 7. Table 6 can show that the MR-WER is going high when we ask for more than one occurrence in the reference for correct word. It is also notable in the case of five references, when the algorithms ask for at least two or three counts for the correct word, the MR-WER is 65.5%, and 77.5% respectively compared to 80.8% average WER in the case of single reference. This is an evidence that while asking for more than one proof in the reference for each correct word, the MR-WER is still outperforming the standard WER when we average it over five references.

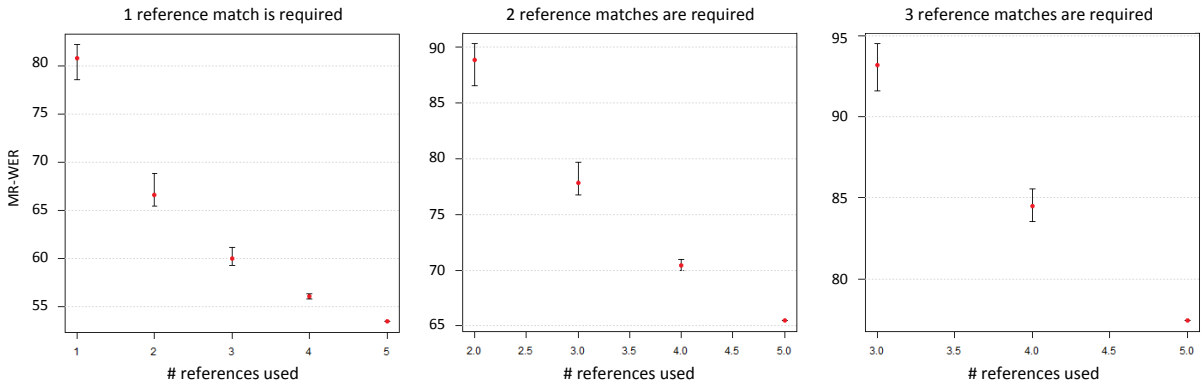

Figure 3: MR-WER for counting correct once or more

	One	Two	Three	Four	Five
1+	80.8%	66.7%	60.0%	56.0%	53.5%
2+	NA	88.8%	77.8%	70.4%	65.5%
3+	NA	NA	NA	84.5%	77.5%

Table 6: MR-WER for counting correct once or more.

6 Conclusion

In this paper, we have presented an innovative way for measuring the accuracy for speech recognition system in non-standard orthographic language; Multi-Reference Word Error Rate (MR-WER). Figure 3 summarized our findings in the multi reference approach applied on Dialectal Arabic (DA). We were able to report 53% MR-WER for five references collectively, while for the same test set the standard WER was between 76.4% to 80.9% when it used the same five references individually. We plan to extend this work to learn from multiple transcription the best orthography to improve the robustness of the computational models. Also, the usage of multi-reference in tuning, and training, similar to the proposed usage in evaluation.

References

[Al-Sabbagh and Girju2012] Rania Al-Sabbagh and Roxana Girju. 2012. Yadac: Yet another dialectal arabic corpus. In *LREC*, pages 2882–2889.

[Ali et al.2014a] Ahmed Ali, Hamdy Mubarak, and Stephan Vogel. 2014a. Advances in dialectal arabic speech recognition: A study using twitter to improve egyptian asr. In *International Workshop on Spoken Language Translation (IWSLT 2014)*, pages http–workshop2014.

[Ali et al.2014b] Ahmed Ali, Yifan Zhang, Patrick Cardinal, Najim Dahak, Stephan Vogel, and Jim Glass. 2014b. A complete kaldi recipe for building arabic speech recognition systems. In *Spoken Language Technology Workshop (SLT), 2014 IEEE*.

[Cardinal et al.2014] Patrick Cardinal, Ahmed Ali, Dehak, Najim, Yu Zhang, Al Hanai, Tuka, Yifan Zhang, James Glass, and Stephan Vogel. 2014. Recent advances in asr applied to an arabic transcription system for al-jazeera. In *INTERSPEECH*.

[Cotterell and Callison-Burch2014] Ryan Cotterell and Chris Callison-Burch. 2014. A multi-dialect, multi-genre corpus of informal written arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*.

[Darwish and Magdy2014] Kareem Darwish and Walid Magdy. 2014. Arabic information retrieval. *Foundations and Trends in Information Retrieval*, 7(4):239–342.

[Diehl et al.2012] Frank Diehl, Mark JF Gales, Marcus Tomalin, and Philip C Woodland. 2012. Morphological decomposition in Arabic ASR systems. *Computer Speech & Language*, 26(4):229–243.

[Elmahdy et al.2014] Mohamed Elmahdy, Mark Hasegawa-Johnson, and Eiman Mustafawi. 2014. Development of a tv broadcasts speech recognition system for qatari arabic.

[Habash et al.2012] Nizar Habash, Mona T Diab, and Owen Rambow. 2012. Conventional orthography for dialectal arabic. In *LREC*, pages 711–718.

[He et al.2011] Xiaodong He, Li Deng, and Alex Acero. 2011. Why word error rate is not a good metric for speech recognizer training for the speech translation task? In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 5632–5635. IEEE.

[Holes2004] Clive Holes. 2004. *Modern Arabic: Structures, functions, and varieties*. Georgetown University Press.

[Mangu et al.2011] Lidia Mangu, Hong-Kwang Kuo, Stephen Chu, Brian Kingsbury, George Saon, Hagen Soltau, and Fadi Biadsy. 2011. The ibm 2011 gale arabic speech transcription system. In *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*, pages 272–277. IEEE.

[Papineni et al.2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.