MULTI-REFERENCE WER FOR EVALUATING ASR FOR LANGUAGES WITH NO ORTHOGRAPHIC RULES

Ahmed Ali[1,2], *Walid Magdy* [1], *Peter Bell*[2], *Steve Renals*[2]

[1]Qatar Computing Research Institute, HBKU, Doha, Qatar
[2]Centre for Speech Technology Research,School of Informatics
University of Edinburgh, Edinburgh EH8 9AB, UK
{amali, wmagdy}@qf.org.qa,{peter.bell, s.renals}@ed.ac.uk

ABSTRACT

Languages with no standard orthographic representation faces a challenge to evaluate the output from Automatic Speech Recognition (ASR). Since the reference transcription text can vary widely from one user to another. We propose an innovative approach for evaluating 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. Results are reported using two Dialectal Arabic (DA) as a language with no standard orthographic; Egyptian, and North African speech. The average WER for the five references individually is 71.4%, and 80.1% respectively. When considering all references combined, the Multi-References MR-WER was found to be 39.7%, and 45.9% respectively.

Index Terms— Under-Resource, WER

1. INTRODUCTION

Word Error Rate (WER) has continued to be the most commonly used metric for evaluating Automatic Speech Recognition (ASR). The metric simply relies on comparing the recognized text to a reference of a manual transcription to the speech signal. This approach has always been seen as sufficient for an effective evaluation of ASR, since transcription of the speech signal is deterministic and one manual transcription should be a sufficient reference. However, in recent years, some interest has been directed towards ASR for dialects and rural languages [1, 2]. Some of these languages suffer from the absence of unified orthographic rules, Non Standard Orthographic Languages (NSO-L). Dialectal Arabic (DA) is an example for NSO-L. Although DA is not a rural language as it is spoken by 300 million people, there is no unique writing system for it. This creates a challenge for evaluating an ASR output, since one reference transcription may cover only a few of many valid forms of the spoken words.

Unlike English, where *enough* is correct word and *enuf* is an incorrect spelling, NSO-L can have many valid written forms for the same word. Table 1 has an example of Egyptian DA in both Arabic and Buckwalter [1], the Table highlights the variations when writing a NSO-L:

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 1: Sample of phrases with multiple valid spellings in Arabic and Buckwalter

In this paper we propose an evaluation methodology for ASR, which accepts the presence of multiple transcription references. The methodology is inspired by the evaluation of Machine Translation (MT) systems, where multiple translation references could be used. Similarly for some languages, multiple spelling and forms could be accepted as a transcription for a word or a phrase. We introduce *multi-reference* WER (MR-WER), which is a modified version of WER that uses multiple reference transcriptions. We describe the process of aligning the multiple references (that can be of differ-

[1]http://www.qamus.org/transliteration.htm

ent lengths); and we show how MR-WER is calculated. We examine our new metric over two different datasets of DA, namely, Egyptian and North African Arabic, that both have no standardized orthography. ubmitteedFor each dialect, we collected a set of five different transcriptions using a crowdsourcing platform, and compared the performance of WER to MR-WER for these dialects.

We provide our scripts and code for calculating MR-WER for the research community for usage and potential future contributions [2]

2. ASR FOR NSO-L

Several studies have investigated applying ASR to under-resourced languages [2]. Under-resourced languages are those lacking the basic components to have a decent ASR system, such as enough labeled speech data for training, a lexicon, and a Natural Language Processing (NLP) pipeline for phonetic systems. Moreover, they can be NSO-L.

DA is considered one of the largest under-resourced languages that is highly used by millions of people in daily conversation and in social media, while lacking most of the required resources for creating an effective ASR. There are many varieties of DDA distributed over the 22 Arabic-speaking countries. Researchers have aggregated DA into four broad regionally-defined language groups: Egyptian, Maghrebi (North African), Gulf (Arabian Peninsula), and Levantine [3, 4]. There are no orthographic rules for writing any of these dialects, which creates large variations in writing, especially on social media websites [5, 6].

In a study by [7], they presented Conventional Orthography for Dialectal Arabic (CODA), explaining the design principles of CODA, and using the Egyptian dialect as an example, which has been presented mainly for the purpose of developing DA computational models. Similar work by [8] studied the best practices for writing Egyptian orthography. They released guidelines for transcribing Egyptian speech for what is called augmented Conventional Orthography for Dialectal Arabic (augmented-CODA). They also reported a gain in Egyptian speech recognition when augmented-CODA is followed in transcribing Egyptian speech data.

In our work, we propose a more robust solution for handling the variations in orthography when no rules exist by using multiple reference transcriptions for evaluations. This leads to less-biased evaluation to a given form of writing. In addition, it is a language independent approach that could be applied to any NSO-L. This has been the main motivation for us not to apply any text normalization or pre-processing for the text.

[2]https://github.com/amali/multiRefWER

3. MULTI-REFERENCE EVALUATION FOR ASR

3.1. Multi-References Alignment to Recognized Speech Text

The initial step for an ASR multi-reference evaluation is to have alignment between each recognized word and the corresponding reference words from all references. Our approach extends 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 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.
- The previous step effectively captures insertions, substitutions, and correct recognitions. However, deletions would not be handled, since there is no corresponding word in the Rec to the deleted words in the reference. In addition, a 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 the 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 references that have nothing deleted would be assigned to "NULL". See Table 2 as an example.

Table 2 shows the output of alignment of a recognized DA sentence with four different references that disagree on the spelling of many words and the number of words itself. As shown, each word in the recognition is aligned to N references, which maximizes the likelihood of finding a possible match that is accepted by one of the references.

3.2. 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.

Index	Rec	$Ref1$	$Ref2$	$Ref3$	$Ref4$
(00-1)	<DEL>	NULL	NULL	nEm	NULL
(00-2)	<DEL>	nEm	nEm	nEm	nEm
(01)	>ETY	Ah	Ah	Ah	hw
(02)	b<n	TbyEy	TbyEy	hw	TbyEY
(03)	dA	<n	dA	TbyEy	dA
(04)	>SlA	dp	>SlA	dh	>SlA
(05)	yEny	>SlAF	yEny	ASlA	yEnY
(06)	<HnA	<HnA	>HnA	AHnA	nHn
(07)	fy	fy	fY	fy	fy
(08)	wDE	wDE	wDE	wDE	wDE
(09)	gyr	gyr	gyr	gyr	gyr
(10)	qAnwny	qAnwny	qAnwny	qAnwny	qAnwnY
(11)	bAlmr	bAlmrp	bAlmrp	bAlmrh	bAlmrh
(12)	gyr	gyr	gyr	gyr	gyr
(13)	dstwry	dstwry	dstwry	dstwry	<INS>
(14)	bAlmr	<INS>	<INS>	<INS>	<INS>
(15)	wADH	<INS>	<INS>	<INS>	<INS>
(16)	>h	<INS>	bAlmrp	<INS>	dstwrY
(17)	fyh	bAlmrp	Ah	bAlmrh	bAlmrh
(18)	AnqlAb	wDE	wDE	wDE	wDE
WER	MR:52%	75%	59%	88%	68%

Table 2: Alignment applied between a recognized text (Rec) and four different references

- **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 are 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 have no corresponding word in one of the references is that 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.

MR-WER is calculated using to the following equation:

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

As shown in Table2, the length of the transcription varies from one reference to another which means that the deletion count is different among different transcriptions. The WER per reference ranged between 59% to 88%, which demonstrates the challenge in using a single reference for evaluation. However, the MR-WER words achieved 52%, which is a more realistic measure for this type of orthography.

4. EXPERIMENTATION

Our experiments were done using data from the Arabic Broadcast News domain, from a DA speech corpus of Al Jazeera broadcasts[9]. For the current study, we chose two dialects; Egyptian (EGY), and North African (NOR). For each dialect, we asked for five transcriptions for each speech segments (utterances), with an average length between 4-6 seconds per utterance. EGY had 2087 utterances, totaling 3.6 hours, and NOR had 1088 utterances with 3.1 hours. The data was transcribed using CrowdFlower[3], a crowdsource platform with a large user base in the Arab world. Quality control was performed using the best practices described by [10].

For the Arabic ASR, we used a grapheme-based system using sequential Deep Neural Network for the acoustic modeling as described in [11]. In [11], it was 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 DA in most cases, and grapheme models will outperform phoneme models. Mainly, the NLP pipeline for the phonetic system is not mature enough for DA, and is still facing challenges such as diacritization, and phonetization. The other amusing feature in the grapheme system is a 1:1 ratio between the number of types and the number of pronunciations 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.

4.1. Inter-Reference Agreement

An initial necessary step before evaluating the effectiveness of our evaluation methodology is to measure the degree of the problem. Here we measure the agreement on the transcriptions among different references. We measure the WER between each two references and apply this to all references for all segments. We found that the median WER among different references for EGY 59% and for NOR 78.5%. We also calculated the percentage of exact-match transcriptions among references. The percentage was only 2.2% and 1.3% for EGY and NOR. These values were astonishing to us. We have looked at many examples and determined this is not an issue of quality control during crowdsourcing. This low inter-reference agreement is not due to bad transcription, rather, it is due to the real, valid variation in the transcription. This highlights the severe issue for these languages and confirms that the evaluation of ASR systems with only one reference would be highly biased.

4.2. MR-WER Results

We have evaluated the ASR output using 1 to 5 reference transcriptions. We have used all the combinations between reference transcriptions in cases when $N > 1$ to validate our findings. As shown in Table 3, for every experiment, we report the minimum, maximum and average MR-WER for each

[3]http://www.crowdflower.com

EGY					
# Ref	One	Two	Three	Four	Five
Min.	69.1%	52.3%	45.9%	42.2%	39.70%
Av.	71.4%	53.4%	46.4%	42.3%	
Max.	74.0%	55.1%	47.3%	42.7%	
# Exp.	5	10	10	5	1
NOR					
# Ref	One	Two	Three	Four	Five
Min.	78.9%	59.1%	51.8%	48.1%	45.9%
Av.	80.2%	60.4%	52.8%	48.7%	
Max.	80.7%	62.2%	53.9%	49.2%	
# Exp.	5	10	10	5	1

Table 3: MR-WER for various number of references per experiment

number of transcriptions we use. We conclude from these experiments two findings:

1. The WER reduces considerably when we increase the number of transcriptions, and it 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 MR-WER has reduced the error from 71.4% to 39.7% in EGY, and from 80.1% to 45.9% in NOR. This could be happening due to various ways of writing DA and not due to bad ASR.

2. The variance in WER reduces noticeably when the number of references increase. This is due to the fact that multi-reference is capable of capturing some of the variations in transcription, which makes the reported error rate more robust to actual mistakes.

4.3. Applying Voting with Multi-References

In the standard WER, the algorithm will loop over a single 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 a correct word in any of the references to make the WER look better rather than validating these findings. 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 clearly in Table 4. The proposed MR-WER reports 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.

		One	Two	Three	Four	Five
EGY	1+	71.4%	53.4%	46.4%	42.4%	39.7%
	2+	NA	78.3%	63.3%	55.5%	50.7%
	3+	NA	NA	83.7%	69.6%	61.6%
		One	Two	Three	Four	Five
NOR	1+	80.2%	60.4%	52.8%	48.7%	45.9%
	2+	NA	84.5%	69.7%	61.6%	56.7%
	3+	NA	NA	88.9%	76.0%	67.5%

Table 4: MR-WER with voting.

5. CONCLUSION

We have presented an innovative way for measuring ASR performance in non-standard orthographic languages; Multi-Reference Word Error Rate (MR-WER). Our results were based on two Dialectal Arabic corpora; Egyptian and North African. We were able to report 39.7%, and 45.9% MR-WER respectively using five reference transcriptions collectively, while for the same test set the average WER was 71.4%, and 80.1% respectively 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, we plan to explore the usage of multi-reference in tuning, and training, similar to the proposed usage in evaluation.

6. REFERENCES

[1] Ahmed Ali, Yifan Zhang, and Stephan Vogel, "Qcri advanced transcription system (qats)," in *Spoken Language Technology Workshop (SLT), 2014 IEEE*, 2014.

[2] Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz, "Automatic speech recognition for under-resourced languages: A survey," *Speech Communication*, vol. 56, pp. 85–100, 2014.

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

[4] Rania Al-Sabbagh and Roxana Girju, "Yadac: Yet another dialectal arabic corpus.," in *LREC*, 2012, pp. 2882–2889.

[5] Kareem Darwish, Walid Magdy, and Ahmed Mourad, "Language processing for arabic microblog retrieval," in *Proceedings of the 21st ACM international conference on Information and knowledge management*. ACM, 2012, pp. 2427–2430.

[6] Kareem Darwish and Walid Magdy, "Arabic information retrieval," *Foundations and Trends in Information Retrieval*, vol. 7, no. 4, pp. 239–342, 2014.

[7] Nizar Habash, Mona T Diab, and Owen Rambow, "Conventional orthography for dialectal arabic.," in *LREC*, 2012, pp. 711–718.

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

[9] Samantha Wray and Ahmed Ali, "Crowdsource a little to label a lot: Labeling a speech corpus of dialectal arabic," in *Interspeech*, 2015.

[10] Samantha Wray, Hamdy Mubarak, and Ahmed Ali, "Best Practices for Crowdsourcing Dialectal Arabic Speech Transcription," in *Proceedings of Workshop on Arabic Natural Language Processing*, 2015, (in press).

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