WERd: Using Social Text Spelling Variants for Evaluating Dialectal Speech Recognition

Citation for published version:

Digital Object Identifier (DOI):
10.1109/ASRU.2017.8268928

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2017)

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
We study the problem of evaluating automatic speech recognition (ASR) systems that target dialectal speech input. A major challenge in this case is that the orthography of dialects is typically not standardized. From an ASR evaluation perspective, this means that there is no clear gold standard for the expected output, and several possible outputs could be considered correct according to different human annotators, which makes standard word error rate (WER) inadequate as an evaluation metric. Such a situation is typical for machine translation (MT), and thus we borrow ideas from an MT evaluation metric, namely TERp, an extension of translation error rate which is closely-related to WER. In particular, in the process of comparing a hypothesis to a reference, we make use of spelling variants for words and phrases, which we mine from Twitter in an unsupervised fashion. Our experiments with evaluating ASR output for Egyptian Arabic, and further manual analysis, show that the resulting WERd (i.e., WER for dialects) metric, a variant of TERp, is more adequate than WER for evaluating dialectal ASR.

Index Terms— Automatic speech recognition, dialectal ASR, ASR evaluation, word error rate, multi-reference WER

1. INTRODUCTION

Automatic Speech Recognition (ASR) has shown fast progress recently, thanks to advancements in deep learning. As a result, the best systems for English have achieved a single-digit word error rate (WER) for some conversational tasks \(^1\). However, this is different for dialectal ASR, for which the WER can easily go over 40\(^\%\).\(^2\)

In a standardized language such as English, we know that *enough* is a correct spelling, while *enuf* is not. However, we cannot be sure about the correct spellings of dialectal words; at best, we would know what a preferred or a dominant spelling is. This is because dialects typically do not have an official status and thus their spelling is not regulated, which opens widely the door to orthographic variation.

\(^1\)Note that here we target primarily intra-dialectal variation. Yet, there is also inter-dialectal variation, e.g., between the different dialects of Arabic.

\(^2\)The global mapping file can help for handcrafted variants like *color/colour* and *ten/10* in English. However, it is not applicable to dialectal Arabic, where multiple spelling variants are acceptable; we use 11M pairs.
More importantly, it is hard to evaluate such a system and to measure progress as multiple possible text outputs for the same speech signal could be considered correct by different people. Thus, there is need for an evaluation measure that would allow for common spelling variations. In this work, we propose to mine such variations from dialectal Arabic tweets and to incorporate them as spelling variants as part of a more adequate ASR evaluation measure for dialects.

Previously, the problem was addressed using the multi-reference word error rate (MR-WER) \cite{5}, which is similar to the multi-reference BLEU score \cite{6} used to evaluate Machine Translation (MT). However, obtaining multiple references is expensive. Moreover, it could take many human annotators to get good coverage of the possible orthographic variants of the transcription of a speech recording. Thus, we propose to use a single reference, but to perform matching using spelling variants that could capture some of the variation.

This was applied to MT, e.g., for parameter optimization \cite{7}, where additional synthetic references are generated for tuning purposes, or for phrase-based SMT, where paraphrasing is applied to the source side of the phrase table \cite{8}, of the training bi-text \cite{9}, or both \cite{10,11,12,13}. Paraphrasing has been also used for evaluating text summarization \cite{14}.

More relevant to the present work, in MT evaluation, paraphrasing was applied to the output of an MT system \cite{15}. It was also incorporated in measures such as TERp \cite{16}, which is a translation edit rate metric with paraphrases. Indeed, here we borrow ideas from TERp for dialectal ASR, with a paraphrase table (in our case, a spelling variants table), which we mine automatically from a huge collection of tweets in an unsupervised fashion. Our experiments and our manual analysis show that this is a very promising idea.

Our contributions are as follows: (i) We propose a method for automatically collecting spelling and tokenization variations for dialectal Arabic (and, presumably, other languages and language variants) from Twitter data; (ii) We further incorporate these spelling variants in an evaluation metric, WERd, which is variation of TERp, and we demonstrate its utility for dialectal Arabic ASR. We release the code for that metric, as well as the spelling variants we mined and used in the metric\footnote{https://github.com/qcri/werd} eleven million pairs, which we extracted from a seven-billion words corpus of dialectal Arabic tweets.

2. METHOD

We propose a method for evaluating dialectal ASR, which consists of two steps: (i) collecting a large number of spelling variants, which we mine from social media in an unsupervised manner, and (ii) using these spelling variants, with associated probabilities, into an MT-inspired evaluation measure (together with standard unit-cost word insertions, deletions, and substitutions).

![Fig. 1. Diagram of our pipeline for extracting dialectal spelling variants from Twitter.](https://github.com/qcri/werd)
2.1. Mining Spelling Variants from Social Media

We use social media to mine dialectal spelling variants from a collection of half a billion dialectal Arabic tweets. Our approach is language-independent, scalable, and unsupervised, as it assumes no prior knowledge about the language, its dialects, or the data.

We build a list of pairs of spelling variants with probabilities using the following steps (as shown in Figure 1):

First, we collect Arabic tweets. Then, we normalize hashtags, URLs, emoticons. We further drop Arabic diacritics and elongation, and we reduce letter repetitions to maximum three. Our pipeline is an extension to the previous work done in Arabic language processing for microblogs [17].

Next, we extract all n-grams of lengths 5–8. In each n-gram, we consider the first two and the last two words as a context, and the 1–4 words in the middle as a target for this context. For example, for a 5-gram we will have < L1, L2, t1, R1, R2 >, while for an 8-gram we will have < L1, L2, t1, t2, t3, t4, R1, R2 >, where L1 and R1 represent the left and the right context words, and tj are the target words in the middle (1 ≤ i ≤ 2, 1 ≤ j ≤ 4).

Next, we generate pairs of potential spelling variants for targets that share the same contexts. This is subject to the constraint that the normalized Levenshtein distance between the targets is less than t, measured in characters. We tried values between 0.1 and 0.6 for t, and we manually inspected the resulting pairs of spelling variants. Ultimately, we set t = 0.6. With normalization in mind, we further impose a constraint that in each pair of spelling variants, one of the targets is extracted in the same contexts at least N times more frequently than the other one (we set N to 3). Finally, with each pair of spelling variants, we associate a score: the average of the two Levenshtein distances.

The first column (yellow) contains the frequent form, which is the target mAfY. The second column (green) contains the source mAFAfY, which is a less frequent term. The next column is the frequency of the source in the same context, e.g., the word mAFAfY occurred 752 times. The following column is the frequency of the source in the same context, e.g., mAFAfY occurred 75 times. Finally comes the normalized Levenshtein distance.

2.2. Using the Spelling Variants for Evaluation: WERd

We borrow ideas from an evaluation measure for MT evaluation, namely Translation Edit Rate Plus or TERp [22]. TERp allows block alignment of words, called shifts within the hypothesis as a low cost edit, a cost of 1, the same as the cost for inserting, deleting or substituting a word. TERp uses a greedy search and shift constraints to both reduce the computational complexity and to model the quality of translation better. The metric further supports tuned weights for the edit operations, a paraphrase table, synonym/hyponym-based matching using WordNet, etc.

The main motivation for using paraphrases in TERp for MT evaluation is to capture some lexical variation, e.g., (controversy over, polemic about), (by using power, by force), (brief, short), (response, reaction). In contrast, we focus on capturing spelling variation in a dialect as shown in Table 1.

In this work, we only use the paraphrasing capability of TERp. We restrict the matching to monotonic, i.e., no reorderings and no shifts. The only additional operation that we allow, compared to WER, is mapping between the hypothesis and the reference using a pair of spelling variants from our spelling variants table, which can span up to four words on either side of the pair of spelling variants as we have explained above. This monotonic version of TERp, with no reordering but with spelling variant matching capabilities gives rise to our metric for dialectal ASR evaluation, which we will call WERd (or WER for dialects).

3. EXPERIMENTS AND EVALUATION

3.1. Dialectal Data

Speech Data. We collected two hours of Egyptian Arabic Broadcast news [23] speech data, which we split into 1,217 segments, each 3-10 seconds long. It can be argued that Egyptian Arabic, which is one of the Arabic dialects, is a language with no established orthographic rules. This makes it difficult to develop standard transcription guidelines covering orthography. Therefore, we decided to have multiple transcriptions, but to let transcribers write the transcripts as they deemed correct, while trying to be as verbatim as possible. All the transcribers are native speakers of the chosen dialect with no linguistic background. Table 2 shows the overlap between the annotators, at the segment level, for their original transcription and after applying surface normalization for alef, yah and hah, which is standard for Arabic. In Table 2 the first number is for the original text, and the second number is for the normalized text. We can see that even after normalization, there are about 15% differences between most of the annotators.

The transcribers were asked to follow these transcription guidelines: http://alt.qcri.org/resources/MGB-3/Arabic_Transcription%20_Guidelines_20170330.pdf

3Below, we will report results after normalization only.
**Social Media Data.** We further collected dialectal Arabic tweets in order to extract spelling variants. In particular, we issued queries using `lang:ar` against the Twitter API. Note that we did not try to control the location where the tweets originated from, but only the language they were written in. We collected two months of tweets (from December 2015 and January 2016), with about eight million tweets per day on average, which yielded a total of half a billion tweets containing over seven billion word tokens.

### 3.2. ASR System

For our experiments, we used the speech-to-text transcription system built as part of QCRI’s submission to the 2016 Arabic Multi-Dialect Broadcast Media Recognition (MGB-2) Challenge [24]. Here are some key features of that system:

**Data:** The training data consisted of 1,200 hours of transcribed broadcast speech data collected from the Aljazeera news channel. In addition, we had ten hours of development data [25]. We used data augmentation techniques such as speed and volume perturbation, which increased the size of the training data to three times the original size [26].

**Speech lexicon:** We used a grapheme-based lexicon [27] with 900k entries, which we constructed using the words that occurred more than twice in the training transcripts.

**Speech features:** To train the acoustic models, we used 40-dimensional high-resolution Mel Frequency Cepstral Coefficients (MFCCs), extracted for each speech frame, which we concatenated with 100-dimensional i-Vectors per speaker in order to facilitate speaker adaptation [28].

**Acoustic models:** We experimented with three acoustic models: Time-Delayed Neural Networks (TDNNs) [29], Long Short-Term Memory Recurrent Neural Networks (LSTMs), and Bi-directional LSTMs (BiLSTMs) [30]. The latter acoustic model outperforms the other two models in terms of WER. We trained all models using Lattice-Free Maximum Mutual Information (LF-MMI) [31] using the Kaldi toolkit [32].

**Language model:** We trained two n-gram language models (LMs). The first one, a tri-gram LM (KN3) used the spoken utterance transcripts for the 1,200 hours. We used this LM in order to generate decoding lattices. We then rescored these lattices using a four-gram LM (KN4), which we trained on the in-domain data and on some extra text. We used interpolated Kneser-Ney smoothing for both LMs, which we built using the SRILM toolkit [33]. We further trained a Recurrent Neural Network Language Model with MaxEnt Connections (RNNME) using the RNNLM toolkit [34].

**Overall ASR system:** We combined the three aforementioned acoustic models, and for the second pass we additionally used the four-gram and the RNNLM for re-scoring the decoded speech lattices. The overall performance was 14.7% WER on the MGB-2 tasks, and this was the best result achieved at the challenge.

![http://dev.twitter.com/](http://dev.twitter.com/)

<table>
<thead>
<tr>
<th>ref1</th>
<th>ref2</th>
<th>ref3</th>
<th>ref4</th>
<th>ref5</th>
</tr>
</thead>
<tbody>
<tr>
<td>77/86</td>
<td>80/84</td>
<td>78/86</td>
<td>80/87</td>
<td></td>
</tr>
<tr>
<td>74/83</td>
<td>71/85</td>
<td>72/85</td>
<td></td>
<td></td>
</tr>
<tr>
<td>77/84</td>
<td>78/84</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>91/93</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 2:** Pairwise overlap of the five human references before/after normalization (in %).

<table>
<thead>
<tr>
<th>WER</th>
<th>TER</th>
<th>WERd</th>
<th>MR-WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>ref1</td>
<td>46.2</td>
<td>37.4</td>
<td>34.3</td>
</tr>
<tr>
<td>ref2</td>
<td>42.9</td>
<td>38.7</td>
<td>35.7</td>
</tr>
<tr>
<td>ref3</td>
<td>48.9</td>
<td>41.9</td>
<td>38.3</td>
</tr>
<tr>
<td>ref4</td>
<td>46.2</td>
<td>39.0</td>
<td>35.6</td>
</tr>
<tr>
<td>ref5</td>
<td>46.0</td>
<td>38.3</td>
<td>34.9</td>
</tr>
<tr>
<td>ALL refs</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

**Table 3:** WER vs. TER vs. WERd vs. MR-WER, after normalization (in %).

### 3.3. Experimental Results

We first evaluated the ASR system on our two-hour dialectal Arabic test dataset using WER with respect to each of the five references. The results are shown in Table 3. We can see that the WER is much higher on our dialectal Arabic dataset, ranging in 40–50%.

We further calculated MR-WER for our ASR system using all five references, achieving a score of 25.3%. This number is much lower than when evaluating with respect to any individual reference, which is to be expected, as we allow more matching options.

Table 3 reports TER and WERd scores calculated with respect to each of the five references and shows for both metrics a strong correlation with WER. We can also see that the scores for WERd are halfway between WER and MR-WER (e.g., for ref1, it is 34.3 vs. 46.2 and 25.3, respectively), but without the need for additional human references.

There are two reasons for MR-WER to be considerably lower compared to the other metrics. First, the way foreign words and code-switching are handled by different annotators, e.g., words like BBC can be written in either Arabic (BBC) or Latin characters. Some annotators would use Arabic while other would prefer English, which would allow matching either of them when evaluating the ASR output with multiple references. Second, in dialectal Arabic, there are many filler words such as أصل which some annotators would skip and some would keep.
Table 4: WERd using pairs of spelling variants extracted using different maximum edit distances (ED).

<table>
<thead>
<tr>
<th>ref</th>
<th>ED ≤ 0.1</th>
<th>ED ≤ 0.2</th>
<th>ED ≤ 0.3</th>
<th>ED ≤ 0.4</th>
<th>ED ≤ 0.5</th>
<th>ED ≤ 0.6</th>
</tr>
</thead>
<tbody>
<tr>
<td>ref1</td>
<td>37.4</td>
<td>37.1</td>
<td>36.6</td>
<td>35.5</td>
<td>34.8</td>
<td>34.6</td>
</tr>
<tr>
<td>ref2</td>
<td>38.7</td>
<td>38.4</td>
<td>37.9</td>
<td>36.9</td>
<td>36.3</td>
<td>36.1</td>
</tr>
<tr>
<td>ref3</td>
<td>41.9</td>
<td>41.5</td>
<td>40.9</td>
<td>39.7</td>
<td>38.9</td>
<td>38.7</td>
</tr>
<tr>
<td>ref4</td>
<td>39.0</td>
<td>38.6</td>
<td>38.1</td>
<td>36.9</td>
<td>36.2</td>
<td>36.0</td>
</tr>
<tr>
<td>ref5</td>
<td>38.3</td>
<td>37.9</td>
<td>37.3</td>
<td>36.2</td>
<td>35.6</td>
<td>35.3</td>
</tr>
</tbody>
</table>

4. DISCUSSION

WERd for different thresholds. Table 4 shows the performance of WERd when using pairs of spelling variants with different maximum edit distances: 0.1–0.6. As the threshold increases, WERd decreases, e.g., for ref1, it goes from 37.4 to 34.3, or 8% relative reduction. The difference is due to the number of matched spelling variant pairs, e.g., 865 for ref1.

Analysis of the pairs of spelling variant matches. Next, we study the relationship between the threshold on the maximum edit distance vs. the spelling variant table size, the number of spelling variants matches, and the accuracy of these matches. This is shown in Figure 3, where we focus the best reference, ref1 (according to native speakers of Egyptian Arabic who have a linguistic background). We can see that the threshold has a major impact on the spelling variant table size: going from 0.1 to 0.6 yields a six times larger table. It also yields a 21 times larger number of spelling variant matches on the test dataset: from 41 to 865.

Of course, this comes at a cost: while all 41 spelling variant matches at threshold of 0.1 are correct, there are 8% errors among the 865 matches at threshold of 0.6. We believe this is a relatively small price to pay, given the advantage of being able to identify 791 additional correct matches, which we capture without the need for having multiple references.

Pearson correlation. We further measured the correlation between WER/MR-WER vs. TER/WERd. We first calculated the scores for WER/MR-WER/TER/WERd for each of the 1,217 test utterances in isolation, and then we calculated the Pearson correlation using the corresponding lists of utterance-level scores. The results are shown in Table 5. We can see that WERd correlates better than TER with both WER and also with MR-WER. Overall, we can conclude that WERd is a promising measure for evaluating ASR systems that target dialectal speech input.

7We were unable to measure the recall as it requires manual evaluation of all the possible candidates for spelling variants in the references.
Closer look at the spelling variants used in test. Finally, we had a closer look at the 865 spelling variant pairs that were matched and used when calculating WERd for the test set of 1,217 segments, when using edit distance of 0.6. Our analysis shows three types of word-level changes:

1. **Word splitting**: 3% of the pairs
   e.g., مقيش -> (mA fy$).

2. **Word merging**: 16% of the pairs
   e.g., زي ما حنا -> (zy mA HnA).

3. **Word substitution**: 81% of the pairs
   e.g., (AlAmyrkAn) -> (الاميركيان).

These statistics show that we learn many useful spelling variants, i.e., more than 80%, rather than just splitting and merging words. Moreover, note that these word-level substitutions are actually small character-level transformations inside words. Tables 6 and 7 show some examples of correct and wrong spelling variant pairs that were matched when calculating WERd for our Dialectal Arabic test set.

Table 8 further shows how spelling variant pairs affect hypothesis scoring for an example test sentence. There are three spelling variant pairs that match the input ASR hypothesis, and finally 3. The former involves word splitting, while the latter two are about substitution. We can see that the WER after using spelling variant substitutions would go down to 30%, (the actual WERd score would be slightly higher as it needs to take the cost of the spelling variant substitutions into account), while the initial WER was 61.5%.

5. **CONCLUSION AND FUTURE WORK**

We have addressed the evaluation of ASR systems that target dialectal speech input, where a major problem is the natural variation in spelling due to the unofficial status and the lack of standardization of the orthography. We have proposed a new metric, WERd (or WER for dialects), a variation of TERp, for which multiple text outputs for the same speech signal can be acceptable given a single reference transcript. Our implementation of WERd was based on mining 11M pairs of spelling variants from a huge dialectal Arabic tweet collection. Our automatic experiments, as well as manual analysis, have shown that this is a highly promising metric that addresses the problems of WER for dialectal speech, and approaches the performance of multi-reference WER.

In future work, we plan experiments with other dialects and non-standardized language varieties. We also want to incorporate word embeddings in the process of computation, e.g., character-based, which can naturally tolerate some spelling variation [35]. We further want to explore using weighted finite state transducers as an alternative way to allow using multiple spelling variants for both references and hypotheses.

6. **ACKNOWLEDGEMENTS**

This work was partially supported by the EU H2020 project SUMMA, under grant agreement 688139.
7. REFERENCES


