PolyWER: A Holistic Evaluation Framework for Code-Switched Speech Recognition

Karima Kadaoui Maryam Al Ali Hawau Olamide Toyin
Ibrahim Ali Mohammed Hanan Aldarmaki
Mohamed bin Zayed University of Artificial Intelligence
{karima.kadaoui, hanan.aldarmaki}@mbzuai.ac.ae

Abstract

Code-switching in speech, particularly between languages that use different scripts, can potentially be correctly transcribed in various forms, including different ways of transliteration of the embedded language into the matrix language script. Traditional methods for measuring accuracy, such as Word Error Rate (WER), are too strict to address this challenge. In this paper, we introduce PolyWER, a proposed framework for evaluating speech recognition systems to handle language-mixing. PolyWER accepts transcriptions of code-mixed segments in different forms, including transliterations and translations. We demonstrate the algorithms use cases through detailed examples, and evaluate it against human judgement. To enable the use of this metric, we appended the annotations of a publicly available Arabic-English code-switched dataset with transliterations and translations of code-mixed speech. We also utilize these additional annotations for fine-tuning ASR models and compare their performance using PolyWER. In addition to our main finding on PolyWER's effectiveness, our experiments show that alternative annotations could be more effective for fine-tuning monolingual ASR models. Our algorithm and additional annotations can be found in our Github repo[1].

1 Introduction

Code Switching (CS) refers to the common phenomenon of mixing two or more languages within a single conversation. When the language switch happens within the same sentence or utterance, it is called intrasentential CS. While language switches between utterances (intersentential) are relatively easy to detect using language identification models, intrasentential code-switching is far more challenging. In Automatic Speech Recognition (ASR), intrasentential CS is particularly difficult to detect and accurately transcribe. Moreover, evaluating the performance of ASR systems in the presence of CS can be tricky due to script differences, transliteration, or non-standard spelling of foreign words.

Our exploration of multilingual pre-trained ASR models such as Whisper (Radford et al., 2023) and Massively Multilingual Speech (MMS) (Pratap et al., 2023b) reveals inconsistencies in CS transcriptions, such as transcribing in the source script, transliterating into the target script, or even translating into the target language. Traditional ASR metrics, namely Word Error Rate (WER) and Character Error Rate (CER), are intolerant to such variations, which can affect the ability to accurately compare the performance of various models. An example of this phenomenon is presented in Table 1, demonstrating how different models exhibit different behaviors in transcribing code-switched speech. In this example, WER treats both CS outputs as incorrect. Yet this evaluation is misleading as both are in fact correct: the first is a translation, the second is a verbatim transliteration.

In this work, we propose a novel variant of the WER algorithm designed to address the shortcomings outlined above. In particular, the algorithm allows the specification of different variants for each word in the reference transcription, including transliteration and translation, resulting in a more tolerant treatment of the variations in CS transcriptions for languages of different scripts. For transliteration, we use CER to account for the nonstandard spelling of transliterated words; for translation, we utilize a BERT model and the cosine distance metric to match possible translations. To make the algorithm consistent with the logic of WER, which ensures that the output is in fact a correct transcription, we apply tight cutoff points for CER and cosine distance. While the algorithm is flexible to account for all kinds of spelling variations, we implement it specifically for handling predictable variations that arise in CS transcription. In light of this flexibility to handle multiple languages, we

[1] https://github.com/mbzuai-nlp/PolyWER

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 6144–6153

refer to the algorithm as **PolyWER** [2].

While the proposed PolyWER algorithm provides the needed flexibility, we lack CS data sets that include these variations in their reference annotations. To enable the application of this method, we manually annotate one of the available CS datasets to provide additional variants. Namely, we selected the Mixat corpus (Al Ali and Aldarmaki, 2024), which consists of ∼15 hours of speech, roughly half of which includes Arabic-English code-switching. The original annotations use the Arabic and English scripts, making it easy to identify code-switching points in the text. We hire native speakers to provide transliterated and translated annotations for the CS segments. Using this data, we evaluate three large pre-trained ASR models that support the Arabic language: Whisper (Radford et al., 2023), MMS (Pratap et al., 2023b), and ArTST (Toyin et al., 2023). As reported in Al Ali and Aldarmaki (2024), while these models support the Arabic language more generally, their performance on this dataset is very poor for two reasons: the speech is in Emirati Arabic, which is a low-resource dialect, and these models were not trained to transcribe in this variant; the performance degrades even more as a result of code-switching. Yet, some observed translated and transliterated outputs, as shown in Table 1, demonstrate the potential of these models if fine-tuned on the target variety. To test that, we fine-tune each of these models on the training segment of Mixat using the three different transcription varieties, and report their performance using various metrics. In addition, we conduct human evaluations on a smaller set of examples to evaluate the different metrics in their consistency with human judgements. Our results indicate that PolyWER is more consistent with human judgement when compared against standard WER and CER metrics as well as a previously proposed multi-reference WER algorithm, especially on cases where the ASR hypothesis includes a combination of code-switched transcription and transliteration.

GT	يعني الحين قطعت مشوار ف Do you go back?
LAT	لا انت قطعت مشوار فهل تعود الآن؟
GT	يتكلم يقول أنا ابي اتعلم video editing
LIT, LAT	أتمنى أن أتعلم فيديو إيديتين

Table 1: Examples of ASR predictions from Whisper-large-v2 zero-shot that include translations and transliterations. GT: Ground truth. LAT.: Translation. LIT.: Transliteration

2 Related Work

Evaluation Metrics for Code-Switching. Most works on code-switching in ASR still rely on standard ASR metrics like Word Error Rate (WER) and Character Error Rate (CER). While these metrics are a good approximation for performance quality, they are too strict as they rely on a single reference transcription. Code-swtiching introduces the possibility of large variations in possible transcriptions, leading to some newly devised metrics to address this limitation. In Mandarin-English code-mixing for example, Mixed Error Rate (MER) is used to address the difference in lexical units between English and Mandarin (Vu et al., 2012). Chowdhury et al. designed a large multilingual end-to-end ASR model supporting monolingual (English, french), dialectal Arabic, and code-switching content. They analyzed the effect of inconsistent ASR output that results in the same word being transcribed using different writing systems on WER. They handle this by benchmarking the code-switched (CS) ASR results with transliterated WER, where they transliterate the English and French recognized tokens into Arabic script to help disambiguate code-switching errors introduced by the multilingual writing systems supported by ASR. They also created a simple Global Mapping File to transliterate between these languages. Ali et al. introduced multi-reference WER (mrWER), an evaluation methodology for ASR for languages without orthographic rules. Their method uses multiple transcription references for evaluating recognized speech. They examine their approach with two datasets of Dialectal Arabic: Egyptian and North African Arabic.

Arabic Code-Switching Datasets. With the prevalence of code-switching in Arabic contexts (Sabty et al., 2020) comes the need for Arabic ASR systems to capture this diversity. This can only be made possible, however, through the development of corpora that represent the number of code-switching languages in addition to an already complex multi-dialectal nature. One such dataset, ArzEn (Hamed et al., 2020), contains 12 hours of speech from 40 participants speaking in Egyptian Arabic with English code-switching (and the

[2] poly-is Greek for 'many', signifying the ability of the algorithm to treat many spelling variations; it's also a reference to the word 'polyglot'.

Algorithm 1 PolyWER takes multiple references representing original transcriptions, transliterations, and translations. CER and BERT cosine distance are used for matching transliterations and translations, respectively, using α and β as thresholds. The algorithm can be easily modified to accommodate more spelling variations. The returned value, d, can then be divided by the original transcription length to get the final PolyWER score.

Input:
/* r: list of references;
each word in r[2] is a list */
$r[0] = \{r_{00}, \ldots, r_{0n-1}\}$ /* transcription */
$r[1] = \{r_{10}, \ldots, r_{1n-1}\}$ /* transliteration */
$r[2] = \{\{r_{20}\}, \ldots, \{r_{2n-1}\}\}$ /* translation */
$h = \{h_0, h_1, \ldots, h_{m-1}\}$ /* hypothesis */
α /* max CER threshold */
β /* min BERT cosine threshold */
Output: d

Init. d of size $(n+1) \times (m+1)$
for $i = 1$ to $n+1$ **do**
  $d[i, 0] \leftarrow i$
end for
for $j = 1$ to $m+1$ **do**
  $d[0, j] \leftarrow j$
end for
for $i = 0$ to n **do**
  for $j = 0$ to m **do**
    /* original transcription*/
    if $r[0][i] == h[j]$ **then**
      $cost \leftarrow 0$
    /* transliteration */
    else if $r[1][i]$ **and** CER$(r[1][i], h[j]) \leq \alpha$ **then**
      $cost \leftarrow$CER$(r[1][i], h[j])$
    else
      $cost \leftarrow 1$
    end if
    /* Translation */
    $cost_{tr} \leftarrow 1$
    if $cost > 0$ **and** $r[2][i]$ **then**
      /* find closest word in reference */
      best_sim = max(cos$(w, h[j])$ **for** w in $r[2][i]$
      if best_sim $\geq \beta$ **then**
        /* convert to distance */
        $cost_{tr} \leftarrow 1 -$ best_sim
      end if
      $best_d \leftarrow min(d[i, j], d[i+1, j], d[i, j+1])$
      $cost_{tr} \leftarrow best_d + cost_{tr}$
    end if
    /* optimal path */
    $C_{\text{SUB}} \leftarrow d[i, j] + min(cost, cost_{tr})$
    $C_{\text{INS}} \leftarrow d[i+1, j] + 1$
    $C_{\text{DEL}} \leftarrow d[i, j+1] + 1$
    $d[i+1, j+1] \leftarrow \min(C_{\text{SUB}}, C_{\text{INS}}, C_{\text{DEL}})$
  end for
end for
Return d

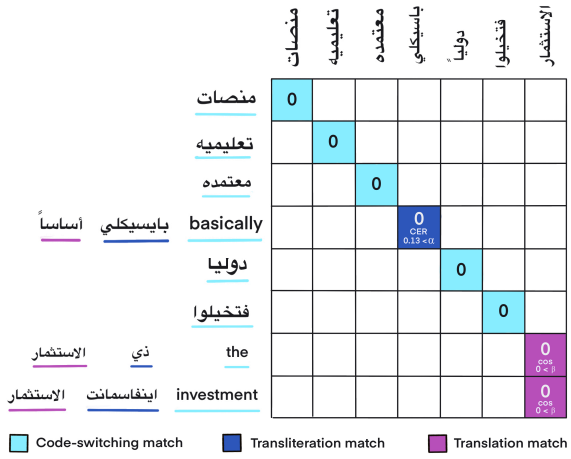

Figure 1: Illustration of the best path in the matrix d for a toy example. The rows and columns correspond to the references and hypothesis, respectively. For brevity, the first row and column of the matrix (used for initialization) are not shown.

occasional French word). Mubarak et al. (2021) introduce QASR, a multi-dialectal speech corpus comprising of 2,000 hours of news speech across 11,092 speakers. It contains 6,000 CS segments of Arabic, French and English although it only represents 0.4% of the dataset. Mixat (Al Ali and Aldarmaki, 2024) is another Arabic CS dataset featuring the Emirati dialect with English mixing. Around 15 hours of podcast recordings are included and 36% of the transcriptions contain code-switching. TunSwitch CS (Abdallah et al., 2024) is a Tunisian dialect dataset with English and French code-mixing thatcarries over 153 hours of unlabeled speech in addition to 10 labeled hours.

3 PolyWER

3.1 Approach

In light of the inconsistent behavior of ASR systems when it comes to language-mixing, we propose a tri-tier approach where edit distance is calculated between words in the hypothesis and three references for each word in the embedded language: original, transliterated, and translated, and the lowest cost is selected. The different costs are computed as follows:

Substitution cost: We perform a traditional comparison between the hypothesis word w_h and reference word w_r where the cost C_{SUB} given our dynamic programming matrix d is expressed as:

$$x = \begin{cases} 0 & w_h == w_r \\ 1 & otherwise \end{cases} \quad (1)$$

$$C_{\text{SUB}} = d[i-1, j-1] + x \quad (2)$$

Transliteration cost: Since transliteration lacks standardized orthography, we are conscious of possible fluctuations in the spelling between the reference and prediction for the same word. We thus use CER instead of exact word match to introduce more lenience in the comparison. We set a hyper-parameter α as a threshold for maximum accepted CER. We express the transliteration cost C_{LIT} as:

$$y = \begin{cases} CER(w_r, w_h) & CER <= \alpha \\ 1 & otherwise \end{cases} \quad (3)$$

$$C_{\text{LIT}} = d[i-1, j-1] + y \quad (4)$$

Translation cost: Since equivalent translations can differ in size, we compute the cosine similarity between the BERT embeddings (Devlin et al., 2019) of the hypothesis word w_h and all the reference words $w_{r_{1:n}}$ in the translated section. Similarly to the transliteration, we set a hyper-parameter β as a threshold for the minimum acceptable similarity value. The cost C_{LAT} is expressed as the complement of the highest cosine similarity:

$$\cos_{\max} = \max_{w_{r_i} \in \text{section}} \cos(\text{BERT}(w_h), \text{BERT}(w_{r_i})) \quad (5)$$

$$z = \begin{cases} 1 - \cos_{\max} & \cos_{\max} >= \beta \\ 1 & otherwise \end{cases} \quad (6)$$

$$C_{\text{LAT}} = \min(d[i-1, j], d[i, j-1], d[i-1, j-1]) + z \quad (7)$$

In equation (7), note that we consider all possible paths, including ones that in regular WER are considered deletions or insertions. The reason is that translations could be aligned between two sequences of varying length, so a diagonal transition (corresponding to matches or substitutions only) would limit this possibility of many-to-one and one-to-many translation.

Final cost: The final cost at each cell in the matrix d is the minimum of the proposed new costs and the traditional insertion and deletion costs:

$$d[i, j] = \min(C_{\text{DEL}}, C_{\text{INS}}, C_{\text{SUB}}, C_{\text{LIT}}, C_{\text{LAT}}) \quad (8)$$

The complete algorithm is described in Algorithm 1, and Figure 1 shows a toy example with the resultant lowest-cost PolyWER alignment.

3.2 Detailed Example

Given a segment that contains code-switching, we have three references r_{CS}, r_{LIT}, and r_{LAT} such as:

r_{CS}: أنا مستقيم في موضوع different تمامًا اللي هو thermodynamics laws

r_{LIT}: أنا مستقيم في موضوع ديفرنت تمامًا اللي هو ثيرموداينااميكس لوز

r_{LAT}: أنا مستقيم في موضوع مختلف تمامًا اللي هو قوانين الديناميات الحرارية

Before we start computing the edit distance, an alignment of the references is required for two reasons: (1) the reference length is used for one of the dimensions of the d matrix, but r_{LAT}'s length can differ from the other two references; and (2) a one-to-one mapping is required between the words in r_{CS} and the equivalent words in r_{LIT} and r_{LAT}. Assume that segments in the embedded language are identified in the references. Given n_i words in a code-switched segment i in r_{CS} and m_i words in the corresponding translated section i in r_{LAT}, we group all m_i words and duplicate them n_i times as shown in Table 2.

Let h be a model's hypothesis:

h: أنا مستقيم في موضوع مختلف تمامًا اللي هو قوانين الثيرموداينااميكس

The hypothesis in this example, which is true in meaning to the audio, contains a mix of transliterations (marked in blue) and translations (marked in purple). While the first four words in h will match exactly with r_{CS}, $h[4]$ will be compared against the other references. Assuming we set α to 0.25 and β to 0.85, the CER between $h[4]$ (مختلف) and $r_{\text{LIT}}[4]$ (ديفرنت) will be too high to pass the α threshold but the cosine similarity between $h[4]$ and $r_{\text{LAT}}[4]$ will be 0 since the words are identical. Similarly, and despite the last two words in h both corresponding to a single code-switching section in the audio, قوانين will match r_{LAT} and الثيرموداينااميكس will match r_{LIT}.

4 Dataset

4.1 Mixat

The Mixat dataset (Al Ali and Aldarmaki, 2024) is an ASR dataset in Emirati dialect code-switched with English. The data was collected from two pod-

Ref.	0	1	2	3	4	5	6	7	8	9
r_{CS}	أنا	مستقيم	في	موضوع	different	تمامًا	اللي	هو	thermodynamics	laws
r_{LIT}	أنا	مستقيم	في	موضوع	ديفرنت	تمامًا	اللي	هو	ثيرموداينامیکس	لوز
r_{LAT}	أنا	مستقيم	في	موضوع	مختلف	تمامًا	اللي	هو	قوانين الديناميات الحرارية	قوانين الديناميات الحرارية

Table 2: Visual representation of the alignment of r_{TR}, r_{LIT}, and r_{LAT} that the PolyWER algorithm expects.

r_{CS}	من البدايه، أنا كنت أعرف هالشي [My passion was architecture]
r_{LIT}	من البدايه، أنا كنت أعرف هالشي [ماي باشون واز أركيتيكتشور]
r_{LAT}	من البدايه، أنا كنت أعرف هالشي [كان شغفي هو الهندسة المعمارية]

Table 3: Example transcription from the Mixat dataset, in addition to our added transliterated and translated versions.

casts where each one represents a split. The train split contains interviews across different episodes between a host and a new guest, and the test split consists of a monologue from a single speaker. The dataset contains approximately 15 hours of speech with 36% of the transcriptions including English code-switching. The Arabic and code-switched parts of the speech were transcribed in Arabic script and latin letters, respectively.

4.2 Expanding on Annotations

Since PolyWER relies on different types of transcriptions that are not all available in Mixat or any other CS dataset, we created two additional copies of the ground-truth labels and replaced the code-switched sections with transliterations in one copy and translations in the other. We hired a native Emirati speaker to complete these annotations. With the translations and transliterations being in Arabic script, we made use of the square brackets [ ] to delimit the boundaries of the code-switched speech sections. A sample sentence from Mixat and our added transcriptions are presented in Table 3. In addition to the additional transcription, the annotator found many errors in the original reference transcriptions and corrected them. The final annotations include corrected references that can be used to benchmark future evaluations on Mixat.

5 Experimental Settings

To effectively test our algorithm, we select three different ASR models and generate predictions on the Mixat dataset under various settings.

5.1 Models

Whisper: Whisper (Radford et al., 2022) is a multi-task speech-to-text system trained in a supervised manner across many languages and tasks, including speech transcription and translation. Whisper can be used off-the-shelf by providing the language id (e.g. `arabic`) and the task (e.g. `transcribe`) for inference. We used the latest whisper version whisper-lg-v3 for optimal performance.

MMS: The Massively Multilingual Speech (MMS) is a multilingual speech-to-text technology spanning thousands of languages (Pratap et al., 2023a). MMS was pre-trained in a supervised manner for ASR across different languages, and the language id can be specified for inference. They use language adapters to optimize the model for different languages.

ArTST: Arabic Text and Speech Transformer (Toyin et al., 2023) is a pre-trained Arabic text and speech transformer, designed with a focus on the Arabic language, and was pre-trained on a thousand hours of Modern Standard Arabic. Unlike Whisper and MMS, ArTST is not a multilingual model, and is not likely to recognize English, but it has been shown to achieve state-of-the-art performance on Arabic ASR and other speech classification tasks, and was show to have some dialectal coverage.

5.2 ASR settings

For zero-shot experiments, we simply used the pre-trained models without modification, following standard tokenization and normalization schemes specified for each model. For fine-tuning experiments, we conducted experiments using the three types of annotations we have: (1) fine-tuning on original transcriptions with code-switched Arabic and English scripts, (2) fine-tuning on the transliterated transcriptions, and (3) fine-tuning using the translated transcriptions. For evaluation, we use the corresponding type of labeling for standard metrics

a	b	$c(0.5f + 0.5m)$		$c * f$		$c * m$	
		Pr(a, b)	**Sp(a, b)**	**Pr(a, b)**	**Sp(a, b)**	**Pr(a, b)**	**Sp(a, b)**
1 - Human	WER	0.649	0.633	0.627	0.612	0.647	0.637
1 - Human	CER	0.816	**0.821**	**0.829**	**0.828**	0.776	0.793
1 - Human	PolyWER	0.810	0.791	0.789	0.766	0.802	0.793
1 - Human	**PolyWER$_f$**	**0.836**	**0.821**	0.826	0.809	**0.817**	0.810
1 - Human	mrWER (Ali et al., 2015)	0.831	0.814	0.815	0.794	**0.817**	0.810
Human	BLEU	0.231	0.477	0.221	0.407	0.233	0.527
Human	BERTScore (F1)	0.829	0.818	0.811	0.798	**0.817**	**0.813**

Table 4: Pearson correlation (Pr) and Spearman rank correlation (Sp) of different ASR metrics against human scores r (or $1 - r$ for error metrics) across three configurations. c: completeness. f: faithfulness. m: meaning.

a	b	$c(0.5f + 0.5m)$		$c * f$		$c * m$	
		Pr(a, b)	**Sp(a, b)**	**Pr(a, b)**	**Sp(a, b)**	**Pr(a, b)**	**Sp(a, b)**
1 - Human	PolyWER	67.21	57.59	63.93	56.68	66.46	55.15
1 - Human	PolyWER$_f$	**82.55**	**77.91**	**78.95**	**76.47**	**81.16**	**75.08**
1 - Human	mrWER	80.18	74.52	78.25	75.51	77.09	68.46

Table 5: Pearson correlation (Pr) and Spearman rank correlation (Sp) of different ASR metrics against human scores calculated only for the zero-shot setting.

like WER/CER or BERT; for PolyWER, we use all three reference transcriptions in all cases.

5.3 Human Evaluation Methodology

We conducted human evaluations to validate the various evaluation metrics on code-switched speech recognition. We extracted ASR predictions from six different systems for 40 randomly selected code-switched utterances from the Mixat test set. Three native Arabic speakers participated in the human evaluation. Each annotator was presented with a speech audio, and a list of ASR transcriptions. They were tasked with ranking each output on three separate dimensions: completeness (c), faithfulness (f), and meaning (m). The ranking was done on a sliding discrete scale from 1 to 6. We re-scaled the values to be in the range [0-1] before combining the ranking for each transcription using the following formula:

$$r = c(0.5f + 0.5m) \tag{9}$$

where we give equal importance to faithfulness and meaning, with completeness acting as a gate that penalizes incomplete predictions. We also compare the metrics with $c * f$ and $c * m$ to evaluate separately how each metric correlates with faithfulness and meaning.

5.4 PolyWER Settings

PolyWER includes hyperparameters that change its behavior. The hyperparameter α denotes the maximum accepted CER score for transliterations, β denotes the minimum accepted cosine similarity between translations. We use $\alpha = .25$ and $\beta = .85$ in our main experiments. Our BERT model of choice in our evaluation is CAMeL-Lab's `bert-base-arabic-camelbert-da`[3].

PolyWER$_f$: Setting β to a value larger than 1 means that PolyWER will not accept translations as correct. We refer to this variant that only accepts original transcriptions and transliterations as PolyWER$_f$, where f stands for *faithfulness*.

6 Results

6.1 Human Evaluation Results

We report the average human ratio score, and the inter-annotator agreement in terms of Inter-Class Correlation Coefficient (ICC) in Table 6. Agreement between annotators ranges from 0.631 to 0.951, with the ArTST model being the most difficult to rate. In Table 7, we report the average human ratio (r) on the subset used for human evaluation, in comparison with various metrics on the same set.

[3] https://huggingface.co/CAMeL-Lab/bert-base-arabic-camelbert-da

System	Fine -tuning	Avg. Ratio %	ICC
Whisper	Zero-shot	61.62	**0.951**
	Transcription	**89.57**	0.867
MMS	Zero-shot	34.18	0.739
	Transcription	56.93	0.845
ArTST	Zero-shot	38.52	0.689
	Transcription	46.42	0.631

Table 6: Human evaluation score (average ratio) and inter-annotator agreement in terms of interclass correlation coefficient (ICC), for each system used in human evaluation. Ratio: $c(0.5f + 0.5m)$

Model	Whisper		MMS		ArTST	
	Zero	FT	Zero	FT	Zero	FT
Avg. Ratio % ↑	61.6	89.6	34.2	56.9	38.5	46.4
WER % ↓	72.6	57.1	89.3	74.9	82.5	76.0
CER % ↓	38.1	16.7	50.4	28.8	43.8	30.7
PolyWER % ↓	37.1	16.8	69.3	49.6	54.3	42.8
PolyWERf % ↓	41.4	16.8	70.1	49.7	56.1	42.9
mrWER % ↓	38.2	15.4	70.2	40.3	54.7	40.6
BLEU ↑	42.3	69.9	7.8	27.7	22.4	37.9
BERTScore ↑	81.3	90.1	77.5	88.4	83.3	78.9

Table 7: Evaluation scores on the human evaluation subset across different metrics in comparison with human ratio. Ratio: $c(0.5f + 0.5m)$

Discussion: We notice from Table 7 that the ranking of the systems for PolyWER aligns with the human evaluation ranking with the exception of ArTST FT. We can see from Table 6 that this model has the lowest inter-annotator agreement, which prompted us to scrutinize its predictions and the resulting PolyWER scores. The utterances that did not align with the human consensus turned out to follow the same pattern: a large number of errors and a much larger number of words. Due to how error-based metrics compute their scores, such cases result in a low error rate, which does not correspond to the way a human would rate the same prediction. For instance, a 40-word sentence with 10 errors would result in a 25% error rate, while human annotators from our observations seem to penalize a prediction by at least half of the overall score when faced with such a high number of errors (i.e. the number of words for a human does not hold the same importance as it does for an error-based metric). Despite this problem affecting WERs and mrWERs, these metrics benefit from their stricter approach (e.g. penalizing transliterations), which so happens to align with human judgement (albeit for different reasons).

Correlation on Zero-Shot models: PolyWER$_f$ stands out as the most consistent metric across various human evaluation configurations, either achieving the highest correlation or being a close second (as seen in Table 4). The other two metrics that also perform well across multiple configurations are CER and mrWER. We notice that CER's high alignment can be attributed to its flexibility with different spellings of the dialectal parts of the utterance (e.g. اتمنى لكم vs اتمنالكم). Other metrics are more rigid on this front and don't account for the variability that is inherent to dialects that lack a standardized orthography. mrWER comes close to PolyWER on configurations that favor meaning, but falls short on predictions that combine transcription and transliteration (which are both faithful references), such as the ones showcased in Table 9. This leads us to compute the correlation on the zero-shot systems only (which are more likely to generate a combination of reference types, and are more accurate representations of models *in the wild*); for these models, we notice a bigger disparity between the two metrics in terms of correlation with human judgement (Table 5).

PolyWER Hyperparameters: We experimented with varying the hyperparameters α and β. The results are shown in Table 8. Note that increasing the tolerance of CER with the α hyperparameter results in improved correlation, whereas increasing the tolerance for translation does not. Upon closer inspection, we find that the BERT model used to evaluate the similarity for translations is generally unreliable.

α	β	Spearman	Pearson
0.15		0.749	0.726
0.20		0.794	0.776
0.25	0.85	0.810	0.791
0.30		0.812	0.793
0.35		**0.813**	**0.794**
	0.75	0.813	**0.795**
	0.80	**0.813**	0.793
0.25	0.85	0.810	0.791
	0.90	**0.813**	**0.795**
	0.95	0.811	0.794

Table 8: Spearman and Pearson correlations on different values of α and β on the human evaluation subset.

Reference	Prediction
my financial situation بخبركم عن حياتي المالية **Basically**	my financial situation بخبركم عن حياتي المالية **بيسيكلي**
إنتو you don't mix in with the crowd لا تحرون	انتو you don't mix in with the crowd لا تحرروا ان
أعماركم **cocktail** ابدا لا	عماركم **كوكتيل** ابدا لا
أسير ال**gym** أشيل الحديد أو أسوي whatever **yoga** whatever i do	اسير **الجم** اشل الحديدة واسوي whatever **يوغا** whatever I do
أنا عندي every morning **ritual** وهذا ال**ritual** it's just positive one	أنا عندي **رتشول** every morning وهذا **الرتشول** it's just a positive one
بتكون هاي الحلقات فقط للمشتركين أو accessible بس حق ال**subscribers** on apple podcasts	وبتكون هاي الحلقات فقط للمشتركين أو accessible بس حق **السبسكرايبرز** on apple podcasts

Table 9: Examples from the test set where PolyWER correlates with human judgement better than other metrics. Transliterations are shown in **red**.

Model	Tr. Sett.	Test Split	PolyWER	PolyWER$_f$	WER	CER
Whisper	Zero-Shot	All	27.79	28.99	29.92	16.70
		CS	33.12	35.35	37.07	23.78
	Transcription	All	**24.61**	**23.81**	**24.83**	**13.64**
		CS	**26.68**	**26.80**	**27.06**	**15.17**
	Transliteration	All	26.48	28.02	33.06	22.53
		CS	30.96	34.39	38.85	27.88
	Translation	All	26.08	29.05	32.55	22.46
		CS	34.63	36.54	38.65	28.30
MMS	Zero-Shot	All	60.19	60.43	61.28	24.92
		CS	63.78	64.22	65.75	30.31
	Transcription	All	47.53	47.63	47.79	**20.55**
		CS	51.02	51.21	51.46	**22.99**
	Transliteration	All	**45.08**	**45.21**	47.25	21.55
		CS	**47.33**	**47.59**	51.33	26.71
	Translation	All	46.02	46.52	47.08	20.66
		CS	49.22	50.15	51.17	25.55
ArTST	Zero-Shot	All	38.12	38.32	39.23	17.13
		CS	42.62	42.99	44.63	22.23
	Transcription	All	26.36	26.44	**26.70**	**11.32**
		CS	30.87	31.03	**31.49**	**14.19**
	Transliteration	All	**25.69**	**25.82**	28.16	13.99
		CS	**29.63**	**29.86**	34.19	19.58
	Translation	All	27.31	27.97	28.53	14.28
		CS	32.52	33.74	34.75	19.93

Table 10: Results of the ASR systems fine-tuned on the three different settings. We report WER, CER and our metric PolyWER. Tr. Sett.: Training Setting. TR: Fine-tuned on Original Transcription. LIT: Fine-tuned on transliteration. LAT: Fine-tuned on translation.

7 ASR fine-tuning & evaluation

In this section, we report the final performance of the various systems trained on Mixat training set (part 1), and evaluated on Mixat testing set (part 2) using PolyWER and other metrics. For each model, we evaluate three variants:

1. Zero-shot performance without fine-tuning.
2. Fine-tuning on the original transcriptions.
3. Fine-tuning on transliterated annotations.
4. Fine-tuning on translated annotations.

The models are evaluated on the same type of annotation when WER, CER, or BERTScore is used. For PolyWER, all three annotations are used as references. The results are shown in Table 10. We find that all models improve with fine-tuning. However, depending on the metric used, MMS and ArTST may perform better when trained on the transliterated or original set. For example, PolyWER ranks the models trained on the transliterated set higher, whereas WER and CER favor the model trained on the original set. On Whisper, all metrics rank the model trained on the original set (with a mix of Arabic and Latin scripts) higher than the other alternatives. This difference in performance trends may be attributed to the inductive bias in Whisper, which already produces high-quality transcriptions and transliterations in zero-shot settings, and can transcribe in both languages, whereas ArTST is a mono-lingual model with limited pre-training on English.

8 Conclusion

In this paper, we introduced PolyWER, a holistic evaluation framework for code-switching in speech recognition. The algorithm accepts multiple references, including original transcriptions, transliterations, and translations. Using special annotations to identify code-switched segments in references, PolyWER applies suitable metrics for the transliterated (CER) and translated segments to maximize flexibility without compromising its integrity as an error metric. To that end, tight thresholds are applied for accepting transliterations based on CER and translations based on BERT cosine similarity. Our evaluation against human judgement shows that a variant of PolyWER correlates well with human scores by balancing faithfulness and meaning preservation. At the same time, it maintains fine-grained discriminative ability, unlike automatic machine translation metrics like WER or BERTScore that are biased towards strict faithfulness or semantic similarity, respectively. The algorithm is flexible as it incorporates hyperparameters to adjust depending on the desired feature. We find that, using the Arabic BERT model used in our evaluation, including translation in the evaluation results in inferior performance. Manual inspection reveals that the semantic similarity scores between Arabic and English in this model are unreliable, so further analysis is needed to demonstrate the potential of including translations with a more reliable cross-lingual similarity model. Our implementation of PolyWER, alongside the additional annotations for Mixat, are publicly available for research.

Limitations

This work is limited by the scarcity of datasets for code-switched ASR, and in particular by the requirement of having multiple references of the specified kind. While we manually annotated one dataset and validated its use for the purposes outlined in the paper, it remains a limitation that only one language-pair was evaluated. We also found the dataset to contain several inaccurate references. We corrected this on the test set to have accurate evaluations, but the problem persists on the train set, which affects the fine-tuned models. Furthermore, the human evaluation scheme devised for measuring the validity of the various metrics may have its own limitations. For instance, we noted large inter-annotator agreement scores for some models. This indicates that even human judgement scores for may not be reliable. Upon close inspection, we also noticed the unreliability of the BERT model used for translations. Finally, while it was outside of the scope of this paper, we believe it would be worthwhile to explore multiple transcription layers (e.g. equivalent spellings/transliterations).

References

Ahmed Amine Ben Abdallah, Ata Kabboudi, Amir Kanoun, and Salah Zaiem. 2024. Leveraging data collection and unsupervised learning for code-switched tunisian arabic automatic speech recognition. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12607–12611.

Maryam Khalifa Al Ali and Hanan Aldarmaki. 2024. Mixat: A data set of bilingual emirati-English speech. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 222–226, Torino, Italia. ELRA and ICCL.

Ahmed Ali, Walid Magdy, Peter Bell, and Steve Renais. 2015. Multi-reference wer for evaluating asr for languages with no orthographic rules. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 576–580.

Shammur Absar Chowdhury, Amir Hussein, Ahmed Abdelali, and Ahmed Ali. 2021. Towards one model to rule all: Multilingual strategy for dialectal code-switching arabic asr. *Preprint*, arXiv:2105.14779.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Injy Hamed, Ngoc Thang Vu, and Slim Abdennadher. 2020. ArzEn: A speech corpus for code-switched Egyptian Arabic-English. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4237–4246, Marseille, France. European Language Resources Association.

Hamdy Mubarak, Amir Hussein, Shammur Absar Chowdhury, and Ahmed Ali. 2021. QASR: QCRI aljazeera speech resource a large scale annotated Arabic speech corpus. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2274–2285, Online. Association for Computational Linguistics.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2023a. Scaling speech technology to 1,000+ languages. *Preprint*, arXiv:2305.13516.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, et al. 2023b. Scaling speech technology to 1,000+ languages. *arXiv preprint arXiv:2305.13516*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.

Caroline Sabty, Mohamed Islam, and Slim Abdennadher. 2020. Contextual embeddings for Arabic-English code-switched data. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 215–225, Barcelona, Spain (Online). Association for Computational Linguistics.

Hawau Toyin, Amirbek Djanibekov, Ajinkya Kulkarni, and Hanan Aldarmaki. 2023. ArTST: Arabic text and speech transformer. In *Proceedings of ArabicNLP 2023*, pages 41–51, Singapore (Hybrid). Association for Computational Linguistics.

Ngoc Thang Vu, Dau-Cheng Lyu, Jochen Weiner, Dominic Telaar, Tim Schlippe, Fabian Blaicher, Eng-Siong Chng, Tanja Schultz, and Haizhou Li. 2012. A first speech recognition system for mandarin-english code-switch conversational speech. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4889–4892. IEEE.