Code-Switching and Multilingual Speech Recognition JSALT 2022

July 13 2022

- WP1 - ASR
- WP2 - CS Generation
- WP3 - Evaluation
- WP4 - Linguistic Aspects of CS

WP1 - ASR

Can we train code-switching ASR systems using only monolingual data?

WP2 - CS Generation

WP4 - Linguistic Aspects of CS

Are the methods being developed in other work packages generalizable?

WP3 - Evaluation

Motivation - Example

Ref	يعنى بقى نتفسح عاوزين لو او برضه عندنا تمارين لو او و two families ال فبنزور weekends mainly family ال
Hyp	يانغ بقى نتفسح عايزين لو أو بوردو عندنا تمارين لولا واو families فا بالنسور الويك أند زمايلي فاعملي
MinEdits	يعني بقى نتفسح عايزين لو أو بردو عندنا تمارين لو واو two families ال فبنزور الويك أندز ماينلي فاميلي

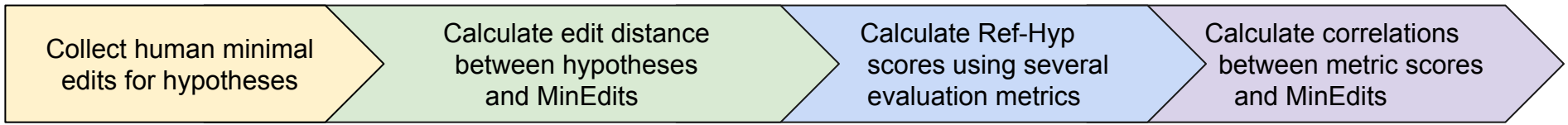
	Hyp-Ref	Hyp-MinEdits
WER	70.0	40.8
CER	47.4	20.0

Matched
Correct and mispenalized
Incorrect but similar phonetically (needs minor edits)
Wrong/missing

Ref	يعني	بقى نتفسح	عاوزين	لو	او	برضه	عندنا تمارين	لو	و او	two	families	ال	فبنزور	ل weekend main family
hyp	يانغ	بقى نتفسح	عايزين	لو	أو	بوردو	عندنا تمارين	لولا	واو		families		فا بالنسو	ويك أند مايلي عملي
minEdit	يعني	بقى نتفسح	عايزين	لو	أو	بردو	عندنا تمارين	لو	واو	two	families	ال	فبنزور	ويك دز ايلي ميلي
CER,														

Work Overview

- Aim:
  - Conduct a study showing how different ASR evaluation metrics correlate with human judgements.
- Overall Plan:



Data Annotation

- Guidelines for Human Minimal Correction (HMC) annotation: https://docs.google.com/document/d/1S_LXwfcFR9gDZlJ0V8Pp-7r1dMm3GEen_hsFuxYNfZI/edit
- 3 Rules:
    1. Rule of Script Segregation:
       Words shall be written in Arabic or Roman script, and not a mix of the two. The only exception is the writing of Arabic affixes and clitics in conjunction with English words. Arabic words should be written in Arabic script and English words can be written in Arabic/Roman script.
       Examples: انجينيرينغ, engineering, exam ال, artiفيشيال

Data Annotation

- Guidelines for Human Minimal Correction (HMC) annotation: https://docs.google.com/document/d/1S_LXwfcFR9gDZlJ0V8Pp-7r1dMm3GEen_hsFuxYNfZI/edit
- 3 Rules:
  1. Rule of Script Segregation:
  2. Rule of Acceptable Readability:
     The transcript should be made readable enough to allow someone to reproduce the original audio and intended meaning. This means there is more than one answer that is acceptable.

عملو	الباتشلر	بتاعهم وعملوا	ماسترز	برا او جوا يعني فعجبتني الفكرة
عملوا	ال bachelor	بتاعهم و عملوا	masters	برا او جوه يعني فعجبتني الفكرة

Data Annotation

- Guidelines for Human Minimal Correction (HMC) annotation: https://docs.google.com/document/d/1S_LXwfcFR9gDZlJ0V8Pp-7r1dMm3GEen_hsFuxYNfZI/edit
- 3 Rules:
  1. Rule of Script Segregation
  2. Rule of Acceptable Readability
  3. Rule of Minimal Edit:
     To correct the ASR, the annotators will do only the most minimal edits on the character-level.

Data Annotation

- Data:
  - Use hypotheses from 3 systems; 1 HMM-DNN and 2 E2E
  - Annotate 2h of speech from ArzEn train set (1.3k sentences) X 3 systems = 3.9k sentences
  - Data annotated by 4 Arabic-English bilingual speakers
  - Inter-annotator Agreement (IAA):
    200/3,900 sentences are annotated by the four annotators
- Languages: Egyptian Arabic-English and Telugu-English

Inter-annotator Agreement (IAA)

CER/WER between AnnotatorX's HMC and hyp (showing amount of edits)

IAA in terms of CER/WER between every 2 annotators

CER/WER between AnnotatorX's HMC and hyp (showing amount of chars/words that would get mispenalized using CER/WER)

	A1	A2	A3	A4	Hyp	Ref
A1		8.3/16.6	6.3/15.0	8.3/18.3	14.9/27.8	23.2/55.4
A2			8.9/17.6	10.8/20.7	14.0/24.5	23.9/56.4
A3				7.0/15.8	15.1/27.9	22.4/54.3
A4					16.0/29.0	24.6/58.5
Average				8.3/17.3	15.0/27.3	23.5/56.2

Disagreement Example

Hyp	يا عصمه manuscript found an أكره
A1	يعني اسمو manuscript found in accra
A2	يعني اسمه manuscript found an أكره
A3	كان اسمه manuscript found in Accra
A4	اسمه manuscript found in accra
Ref	كان اسمه manuscript found in acra

Evaluation Metrics

- Covered metrics:
  - WER
  - MER (Match Error Rate)
  - CER
  - Transliteration
  - Phone edit distance
  - Semantic similarity of translation

Transliteration

		CER	WER
Ref	كان اسمه manuscript found in acra	30.3	66.7
Hyp	ياعصمه manuscript found an أكره		
Tr-En(Ref)	Kan Asmah manuscript found in acra	27.6	66.7
Tr-En(Hyp)	easme manuscript found an Aqrah		
Tr-Ar(Ref)	كان اسمه مانوسكريبت فوند ين اكرا	25.9	66.7
Tr-Ar(Hyp)	ياعصمه مانوسكريبت فوند ان أكره		

Transliteration - Correlation Results

	Transliterating to Arabic		Transliterating to English	
	CER(ref_{TrAR},hyp_{TrAr})	WER(ref_{TrAR},hyp_{TrAR})	CER(ref_{TrEN},hyp_{TrEN})	WER(ref_{TrEN},hyp_{TrEN})
CER(MinEdits,hyp)	0.803	0.372	0.734	0.527
WER(MinEdits,hyp)	0.687	0.441	0.689	0.608

Phone Similarity Edit Distance

- Map the script from the two languages into IPA phones
- Use the phoneme error rate with substitution weight scaled by the similarity between the phones
- Measure the similarity between the phonemes based on the articulation feature vectors: nasal, front, back, labial etc
- Example:
  - Arabic: ا كايند اف
  - English: a kind of
  - Arabic phonetics: a kajnd aof
  - English phonetics: ə kajnd ʌv
    - PER: 0.65
    - PSD: 0.375

		ə	k	a	j	n	d	ʌ	v
a	[0.	1.	2.	3.	4.	5.	6.	7.	8.],
k	[1.	0.113,	1.113,	2.	3.	4.	5.	6.	7.],
a	[2.	1.113,	0.113,	1.113,	2.113,	3.113,	4.113,	5.113,	6.113],
j	[3.	2.113,	1.113,	0.113,	1.113,	2.113,	3.113,	4.113,	5.113],
n	[4.	3.113,	2.113,	1.113,	0.113,	1.113,	2.113,	3.113,	4.113],
d	[5.	4.113,	3.113,	2.113,	1.113,	0.113,	1.113,	2.113,	3.113],
a	[6.	5.113,	4.113,	3.113,	2.113,	1.113,	0.113,	1.113,	2.113],
o	[7.	6.113,	5.113,	4.113,	3.113,	2.113,	1.113,	0.21 ,	1.21],
f	[8.	7.081,	6.113,	5.113,	4.113,	3.113,	2.113,	1.21 ,	0.581],
	[9.	8.081,	7.113,	6.113,	5.113,	4.113,	3.113,	2.21 ,	1.242]]

Phone Similarity Edit Distance

```
ID: 1
REF: a kind of
HYP: ا كايند اوف
REF phone: ə kajnd ʌv
HYP phone: a kaind auf
PER: 0.625 PSD: 0.375 PSD_norm: 0.2258

ID: 2
REF: لا في at least a chance نحاول مرة extra يبقى يبدأ يمكن معاه more flexible
HYP: لا في atlista chance نحاول مرة extra يبقى يبدأ يمكن معي more flexible
REF phone: la fi æt list ə ʧæns nħaul mrt ɛkstɹə mʕah imkn ibda ibqa mɔɹ flɛksəbəl
HYP phone: la fi ætlɪstə ʧæns nħaul mrt ɛkstɹə mʕi imkn ibda ibqa mɔɹ flɛksəbəl
PER: 0.05263 PSD: 0.03112 PSD_norm: 0.02703

ID: 3
REF: artificial
HYP: ارtficial
REF phone: ɑɹtəfɪʃəl
HYP phone: art fɪʃəl
PER: 0.33333 PSD: 0.16844 PSD_norm: 0.0516



Mean PER_tot:      0.14865
 Mean PSD_tot:       0.085
 Mean PSD_norm_tot: 0.05079
```

Phone Similarity Edit Distance - Correlation Results

		weight=2		weight=4		weight=8	
	PER(ref,hyp)	PSD(ref,hyp)	PSD_norm(ref,hyp)	PSD(ref,hyp)	PSD_norm(ref,hyp)	PSD(ref,hyp)	PSD_norm(ref,hyp)
CER(MinEdits,hyp)	0.7434	0.767	0.700	**0.774**	0.734	0.747	0.732
WER(MinEdits,hyp)	0.6478	0.594	0.497	**0.633**	0.544	0.636	0.565

Semantic Similarity Using Translation- Example

		BLEU	chrF	Semantic Similarity
hyp	آ ده أغلب ال أكتفيتيز اللي أنا بعملها في الجامعة يعني			0.7722
ref	اللى انا بعملها فى الجامعة يعنى activities ده اغلب ال			
hyp_ar	آ ده أغلب ال أكتفيتيز اللي أنا بعملها في الجامعة يعني	4.9	49.6	0.8740
ref_ar	أنشطة ده اغلب اللى انا بعملها فى الجامعة يعنى			
hyp_en	Oh, this is most of the activeties that I do at university, I mean	22.9	65.4	0.8665
ref_en	These are most of the activities that I do at the university			
hyp_ja	ああ、これは私が大学で行っている活動のほとんどです、つまり	67.5	82.6	0.9581
ref_ja	これらは私が大学で行っている活動のほとんどです			

Semantic Similarity Using Translation - Correlation Results

	Metric(ref$_{AR}$,hyp$_{AR}$)			Metric(ref$_{EN}$,hyp$_{EN}$)			Metric(ref$_{JA}$,hyp$_{JA}$)			Average		
	BLEU	chrF	Sem	BLEU	chrF	Sem	BLEU	chrF	Sem	BLEU	chrF	Sem
CER(MinEdits,hyp)	0.25	0.55	0.64	0.49	0.62	0.60	0.45	0.49	0.59	0.50	0.65	0.72
WER(MinEdits,hyp)	0.32	0.52	0.59	0.54	0.65	0.61	0.48	0.53	0.53	0.56	0.67	0.68

Results

	CER(ref,hyp)	WER(ref,hyp)	MER(ref,hyp)	Transliteration	PhoneticSimilarity	SemanticSimilarity	Average(Transliteration, PhoneticSimilarity, SemanticSimilarity)
CER(MinEdits,hyp)	0.683	0.372	0.445	0.803	0.774	0.716	**0.820**
WER(MinEdits,hyp)	0.622	0.437	0.516	0.687	0.633	0.680	**0.703**

Discussion

- How do we measure how well a metric reflects human minimal (post-edit) edit effort?
  - CORREL(CER(MinEdit,hyp) vs metricX(ref,hyp))
  - CORREL(WER(MinEdit,hyp) vs metricX(ref,hyp))
  - CORREL(metricX(MinEdit,hyp) vs metricX(ref,hyp))
  - CORREL(WER(MinEdit,hyp) vs metricX(MinEdit,hyp))
- Improve transliteration
- Investigate how we can aggregate scores from different metrics to better correlate with human judgment.

Questions?