Meeting: 05/05/2021

Cécile Macaire

Recap / Fine-tuning wav2vec2 model

Model : `facebook/wav2vec2-large-xlsr-53`

Corpus :

Corpus	Yongning Na	Japhug
Size	≈ 5 hours	≈ 30 hours

Evaluation metrics :

Word Error Rate (WER) / Character Error Rate (CER) / Phoneme Error Rate (PER)

Recap / Fine-tuning wav2vec2 model

Performances on the training data

Language	Model	Train size	WER_huggingface	CER_huggingface
Na	model_60epochs	150 sentences (15 minutes)	0.93921568627451	0.250009224071136
Na	model_60epochs	300 sentences (30 minutes)	0.785490196078431	0.170423938309412
Na	model_60epochs	450 sentences (45 minutes)	0.732549019607843	0.155407150499945
Na	model_60epochs	600 sentences (1h)	0.708627450980392	0.143046895177656
Na	model_60epochs	900 sentences (1h20)	0.667058823529412	0.125816330295539
Na	model_60epochs	1200 sentences (2h)	0.628627450980392	0.13125945606849
Na	model_60epochs	1500 sentences (2h30)	0.61921568627451	0.115908336100963
Na	model_60epochs	1786 sentences (3 h)	0.593333333333333	0.111221816303185
Na	model_60epochs	2100 sentences (3h30)	0.5839215686274509	0.110742093804199
Japhug	model_60epochs	3186 sentences (5h31)		0.115567978880348

Recap / Fine-tuning wav2vec2 model

Performances on the test data

Lang	Model	Tones / Clean ?	Train size	Test size	WER_huggingface	CER_huggingface	PER_difflib	CER_levenshtein	WER_levenshtein
Na	model_60epochs	Tones = 1	1786 sentences (3 h)	336 sentences (44 minutes)	0.490509059534	0.092234824878	0.079223381141	0.07008193383	0.319199182513
Na	model_60epochs	Tones = 0	1786 sentences (3 h)	336 sentences (44 minutes)	0.398619499568594	0.082450008594511		0.059381510518605	0.332173909457734
J	model_60epochs	Keep all test data	3186 sentences (5h31)	350 sentences (44 minutes)	0.239303843364757	0.088794498381877		0.059904970148425	0.257242915932687
J	model_60epochs	Remove data with unmatched timecodes	3187 sentences (5h31)	335 sentences (44 minutes)	0.21890359168242	0.068786045857364		0.046614529768536	0.24356776273533

Recap / Fine-tuning wav2vec2 model - Na results

What are the main errors ?

→ **segmentation errors** on word-level.

Examples:

Ref	['hɑ˧', 'tʰɣ˧ʈɣ˥', 'mɣ˧', 'tʰi˧kʰɯ˩']	length = 4
Pred	['hɑ˧ɖɯ˧ʈɣ˥', 'mɣ˧tʰi˧kʰɯ˩']	length = 2
Ref	['po˧ɖʐɯ˧mɤ˧dʑɤɹhĩɹ', 'ɖɯ˧seɹ', 'tɑ˧qɑ˧', 'tɑ˧qɑ˧', 'tɑ˧qɑ˧', 'pi˧']	length = 6
Pred	['po˧ɖʐɯ˥', 'mɤ˧dʑɤɹhĩɹɖɯɹseɹ', 'tæ˧qæ˧tæ˧qæ˧tɑ˧qɑ˧pi˧']	length = 3

→ **tones** (uni tones = {"˩", "˥", "˧"}, bi tones = {"˩˧", "˧˥", "˥˩", "˥˧"})

Example:

Ref	tɕʰɤ˧ɻni˧ne˧ʝi˥ ʐo˧˥ le˧po˩˧hõ˧ pi˧dʑo˩ zo˧bæ˩ ɖɯ˧ɣ˧ɳɯ˧ mɤ˧ɻni˧ pi˧tsɯ˩mɣ˩ ɑ˩ʁo˧
Pred	tɕʰɤ˧ɻni˧le˧ʝi˩ ʐo˧˥ le˧po˩˧hõ˩ pi˧dʑo˩ zo˧bæ˧ ɖɯ˧ɣ˧ɳɯ˧ mɤ˧ɻni˩ pi˥tsɯ˩mɣ˩ ɑ˩ʁo˧

Recap / Fine-tuning wav2vec2 model - Japhug results

What are the main errors ?

→ **segmentation errors** on word-level.

Example:

Ref	['ki', 'kɯra', 'ɲɯkhama']
Pred	['kɯkɯra', 'nɲɯkhama']

→ **chinese borrowings.**

→ **update transcriptions ?**

→ **wrong timecodes in the transcriptions.**

→ **the external sandhi.**

Example:

Ref	kɤmbɣom mɤra ma aʑo aβlu tu
Pred	kɤmbɣo mara ma ʑo aβlu tu

→ **vowel fusions** (ɯ-ɤ > ɤ, ɤ-ɤ > ɤ, ɯ-a > a, ɤ-a > a, u-o > o).

Example:

Ref	mɯ-ɲɯ-ɤmtɕhoʁ
Pred	mɯɲɤmtɕhoʁ

Current work / N-best hypothesis with beam search

- “synchronous” beam search algorithm.
- generation of the 100-best hypothesis for all the test data.

Id_sentence	Reference	Prediction
0	tʰiʌ|le↑gæ↑1|mɣ↓ɹɯɤ↓tʰɑ↑dzo↓|tʰiʌ|hĩ↑nɯ↓|le↑wo↑1|ə↑mɣ↑kĩ↑|	tʰiʌ le↑gæ↑1mɣ↓ ɹɯɤ↑tʰɑ↑1 dzo↓ tʰiʌ hĩ↑nɯ↓ le↑wo↑ ə↑mɣ↑kĩ↑

Current work / N-best hypothesis with beam search

➢ **Computation of the oracle scores :**

Metrics	WER_HuggingFace	CER_HuggingFace	WER_lev	CER_lev
Oracle score	0.4027450980392157	0.06922764677663383	0.24350706955983692	0.05870925083702372
Gain	8,78 %	2,3 %	7,56 %	1,13 %

Objectives

Short-term

- Generate the predictions with the use of a language model in the decoder.
- Use the n-best hypothesis to improve the predictions.
- Implement a word segmentation model.

Long-term

- Use the knowledge of dictionaries to improve the performances.
- Fine-tune the model on the all japhug corpus.

Open questions - Japhug

- Working on related languages (i.e. rGyalrong languages) ?

- Explore the transfer learning between speakers ?