Exploring Effective Data Utilization For Low-Resource Speech Recognition

Zhikai Zhou, Wei Wang, Wangyou Zhang, Yanmin Qian

X-LANCE Lab, Shanghai Jiao Tong University, China

May 2022

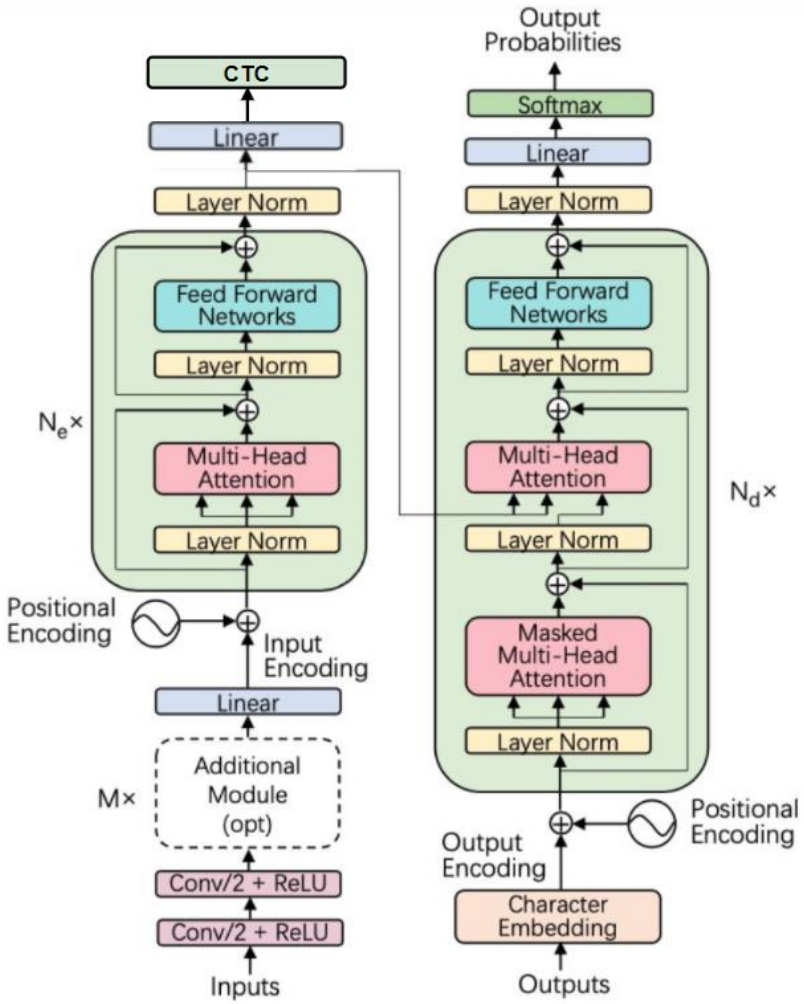
Background

Low-Resource ASR

- There are more than 6,000 languages in the world
- Only several languages have enough data to build the ASR system
- How to build the ASR system for low-resource languages?
  - Data Augmentation
  - Transfer Learning(Utilize data from other languages)
  - Unsupervised/Semi-supervised Learning(Utilize data without label)

Background

Related work – Transformer-based E2E ASR

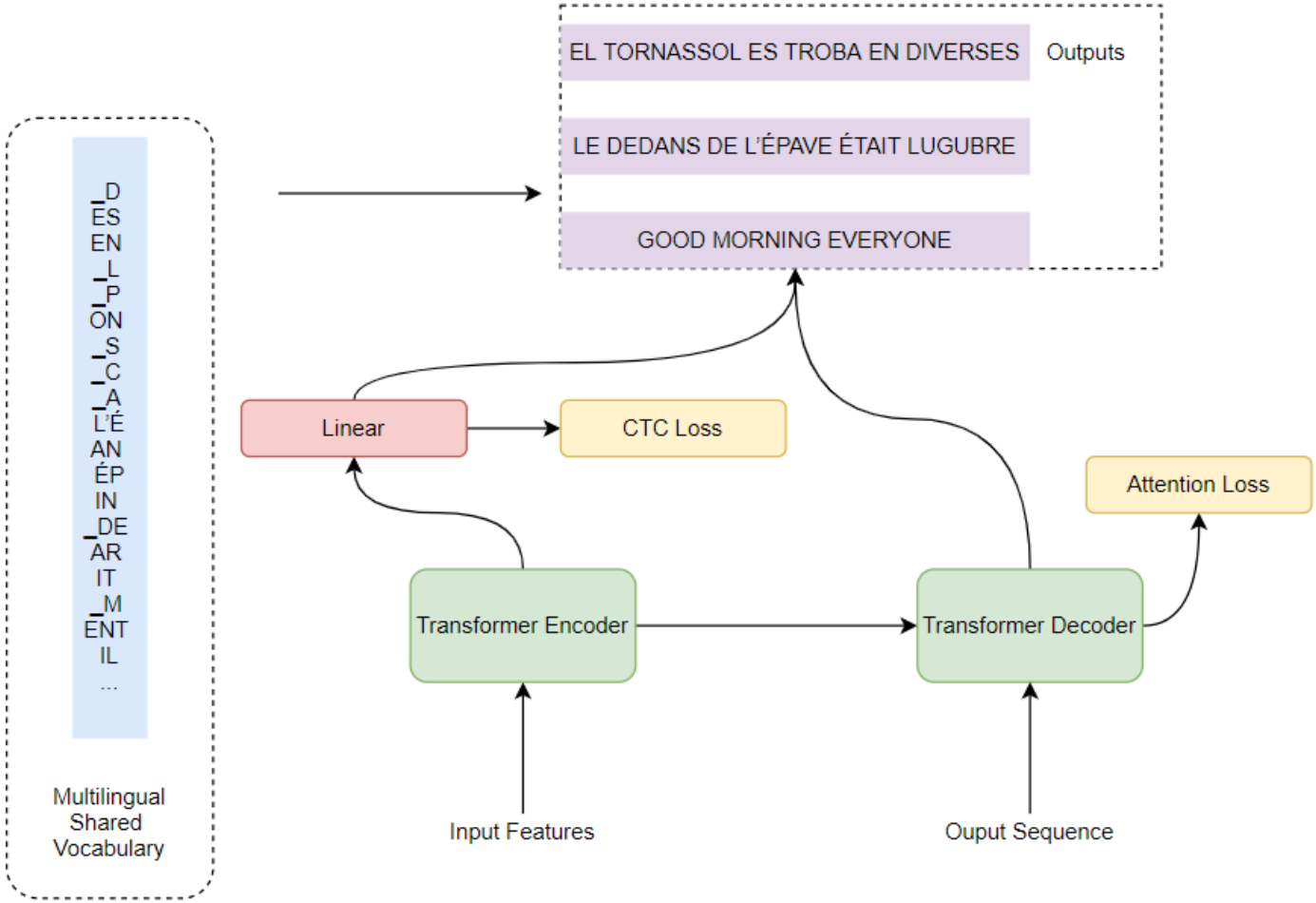


- Encoder transforms features to hidden representations
- Decoder query previous outputs on hidden representations by the attention mechanism
- Model is trained by both the CTC Loss and Cross Entropy.
- The joint decoding is adopted to predict the output sequence.

L. Dong, S. Xu and B. Xu, "Speech-Transformer: A No-Recurrence Sequence-to-Sequence Model for Speech Recognition," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5884-5888, doi: 10.1109/ICASSP.2018.8462506.

Background

Related work – Multilingual Pretraining and Finetuning



- A multilingual ASR model is first pretrained in several rich resource languages
- The vocabulary is shared among all rich resource languages
- Then we finetune the ASR model using the target low-resource language

Picture partly from ‘’Hou W, Dong Y, Zhuang B, et al. Large-Scale End-to-End Multilingual Speech Recognition and Language Identification with Multi-Task Learning[C]//INTERSPEECH. 2020: 1037-1041.”

Proposed Method

- Data Weighing Based on Language Similarity
- Dynamic Curriculum Learning
- Length Perturbation

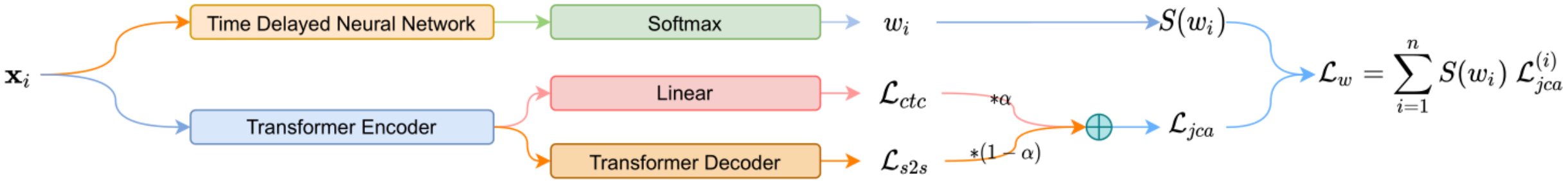
Language	Word	IPA
Catalan	pronunciació	prununsiəsio
French	prononciation	prɔnɔ̃sjasjɔ̃
Italian	pronuncia	pronuntʃa
Portuguese	pronúncia	prunũsja
Basque	ahoskera	aʊʃkerə

IPA: International Phonetic Alphabet

- An example of word “pronunciation” from different languages.
- Basque is very different from others.

Proposed Method

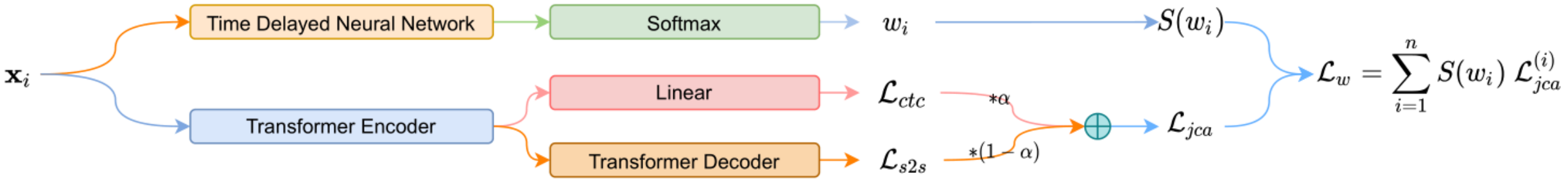
Data Weighing Based on Language Similarity



- To utilize the similarity between the target and non-target languages
- A language classifier is trained to obtain the similarity in utterance level
- The posterior or embedding similarity is used to weigh the utterance in multilingual corpus

Proposed Method

Data Weighing Based on Language Similarity



Posterior based: $w_i = P(y = l|x_i)$

Similarity based:

$$cos\_sim(\mathbf{a},\mathbf{b}) = \frac{\mathbf{a}\cdot\mathbf{b}}{||\mathbf{a}||\,||\mathbf{b}||}$$

$$w_i = \frac{1 + cos\_sim(s_i, \frac{\sum_{k=1}^{|L|} s_k}{|L|})}{2}$$

$$\mathcal{L}_w = \sum_{i=1}^{n} Softmax(w_i)\mathcal{L}_{jca}^{(i)}$$

Sortagrad: Sort samples by lengths

Dynamic Curriculum Learning: Sort samples by Dynamic Difficulty metrics

Difficulty Metrics:

- Loss $$s(\mathbf{x};\theta^t) = \mathcal{L}(\mathbf{x};\theta^t)$$
- Accuracy $$s(\mathbf{x};\theta^t) = -a(\mathbf{x};\theta^t)$$
- Normalized Loss $$s(\mathbf{x};\theta^t) = \frac{\mathcal{L}(\mathbf{x};\theta^t)}{T}$$
- Decline based Loss/Accuracy $$d(\mathbf{x};\theta^t) = -\frac{s(\mathbf{x};\theta^{t-1}) - s(\mathbf{x};\theta^t)}{s(\mathbf{x};\theta^{t-1})}$$

Why Decline based metrics?

- If loss/accuracy has minor changes
  - Low Accuracy: Sample is too hard for model to learn in current status
  - High Accuracy: Sample is too simple for model to learn. No further improvement can be achieved
- Big changes: Model can learn much from this sample

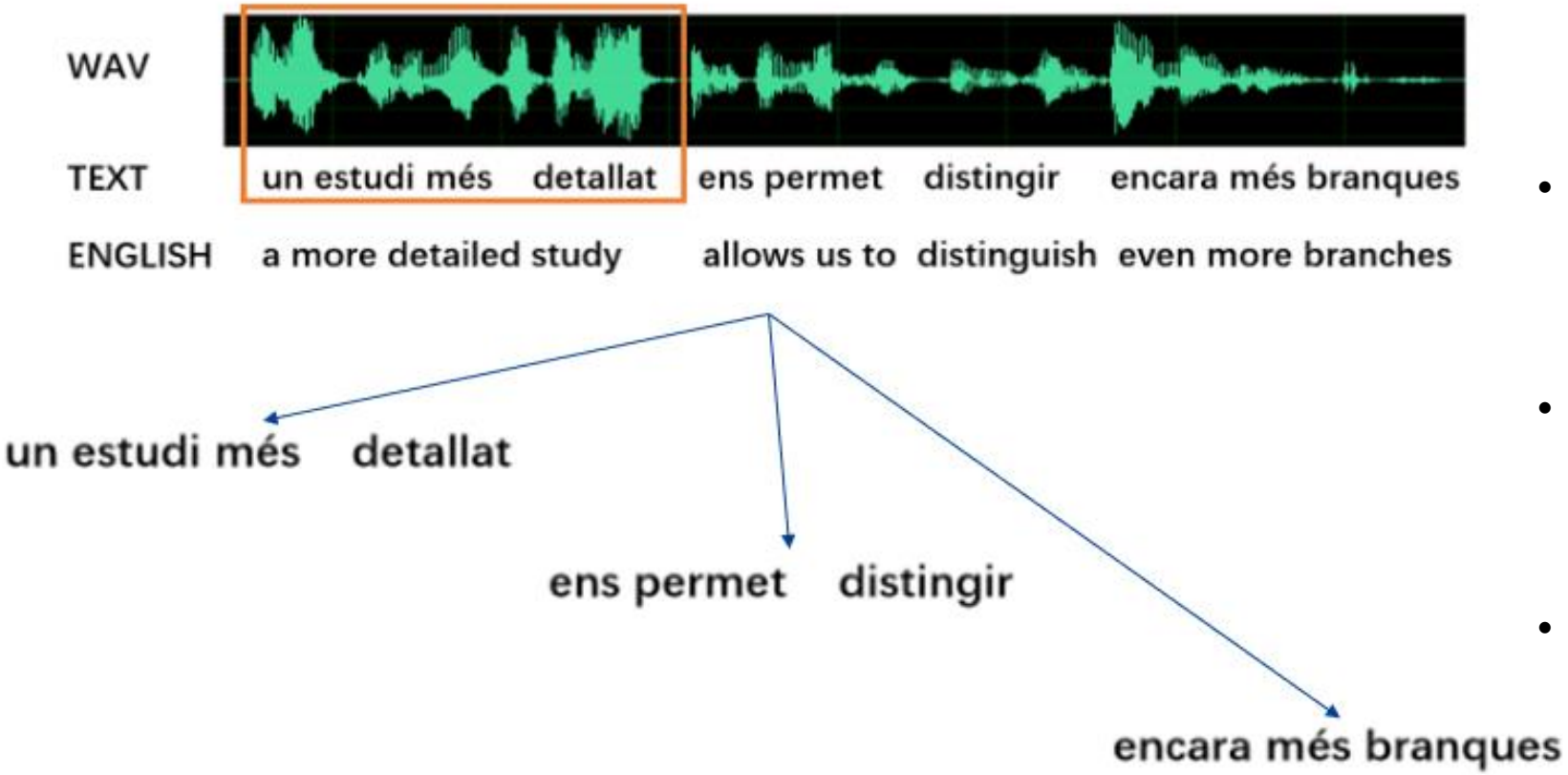
Procedure

- For each phase(several epochs), infer the whole training set to get the loss/accuracy values
- Sort and subset the training set
- Gradually increase the amount of training data to cover the entire training set

$$a(t) = \min\left(1,\ a_0 + \frac{\beta t}{T}(1 - a_0)\right)$$

Proposed Method

Length Perturbation



- Split one utterance into several sub sequences
- Word boundaries are obtained by a hybrid ASR model
- The data can be augmented for several folds.

Common Voice

Language	#Spk.	#Utt.	Duration
Catalan	4,742	317,693	488 hr
Basque	834	61,426	88 hr
Portuguese	717	39,072	48 hr
Italian	4,976	83,407	130 hr
French	11,381	412,332	554 hr

- Rotate the role of the target low-resource language
- Only 10 hours training set for the target language
- Baseline and Speed perturbation (WER) results are as follows

Methods	Catalan dev/test	Basque dev/test	French dev/test	Portuguese dev/test	Italian dev/test
Baseline	21.7/22.0	20.0/21.2	35.8/35.7	19.8/19.0	23.8/23.6
SP_3fold	20.2/20.6	17.5/18.0	32.5/32.5	18.2/17.2	21.3/21.3

Experiment

Results

Method	Catalan dev/test	Basque dev/test	French dev/test	Portuguese dev/test	Italian dev/test
Baseline	21.7/22.0	20.0/21.2	35.8/35.7	19.8/19.0	23.8/23.6
L.Post	21.5/21.7	19.4/19.9	34.6/34.6	19.6/18.6	22.7/22.6
L.Sim	21.5/21.6	19.3/19.9	34.6/34.7	19.4/18.5	22.7/22.7
U.Post	21.2/21.2	19.0/19.5	34.2/34.3	18.0/17.8	22.0/21.8
U.Sim	**20.3/20.4**	**18.5/19.0**	**34.1/34.2**	**17.6/17.0**	**21.2/21.1**

Different data weighing methods(WER%)

- **L/U: Language/Utterance Level**
- **Post: Posterior-based weighing**
- **Sim: Similarity-based weighing**

- The utterance level using the similarity strategy has achieved the best performance.
- All methods achieve better results than the baseline

Experiment

Results

Methods	Catalan dev/test	Basque dev/test	French dev/test	Portuguese dev/test	Italian dev/test
Baseline	21.7/22.0	20.0/21.2	35.8/35.7	19.8/19.0	23.8/23.6
CL	22.0/22.2	20.8/21.7	35.9/35.8	19.9/19.0	23.8/23.7
DCL_A	20.9/21.1	18.8/19.2	34.0/34.1	18.6/17.4	23.0/22.8
DCL_L	21.0/21.0	18.4/19.0	33.6/33.6	18.5/17.4	22.6/22.4
DCL_L*	**20.4/20.6**	**17.4/18.3**	**33.0/33.1**	**17.8/16.7**	**21.8/21.6**

Different curriculum learning methods(WER%)

- **CL: Traditional Sortagrad**
- **DCL_A/L: Dynamic curriculum learning using accuracy/loss**
- **DCL_L*: Dynamic CL using normalized loss**

- DCL using normalized loss achieves the best result

Experiment

Results

Methods	Catalan dev/test	Basque dev/test	French dev/test	Portuguese dev/test	Italian dev/test
Baseline	21.7/22.0	20.0/21.2	35.8/35.7	19.8/19.0	23.8/23.6
SP_3fold	20.2/20.6	**17.5**/18.0	32.5/**32.5**	**18.2/17.2**	21.3/21.3
LP_2fold	20.7/20.6	19.7/20.8	35.9/35.8	19.7/19.0	23.5/23.3
LP_3fold	20.1/20.1	18.7/20.6	34.1/34.1	18.8/18.2	22.1/22.7
LP_4fold	**20.1/19.8**	17.6/**17.9**	**32.4/32.5**	**18.2**/17.5	**20.7/20.6**
LP_5fold	20.2/20.0	18.1/19.1	33.2/33.3	18.7/18.0	21.2/21.0
SP+LP	**18.7/18.8**	**16.8/17.2**	**31.4/31.3**	**17.4/16.3**	**20.7/20.3**

Different Data Augmentation Methods (WER%)

- **SP_3fold: Speed Perturbation(0.9,1.0,1.1)**
- **LP_kfold: k fold length perturbation(1/k,2/k,…,1.0)**
- **SP+LP: SP with LP**

- LP and SP are complementary with each other

Experiment

Results

Integrated Methods(WER%)

Methods	Catalan dev/test	Basque dev/test	French dev/test	Portuguese dev/test	Italian dev/test
M0	21.7/22.0	20.0/21.2	35.8/35.7	19.8/19.0	23.8/23.6
M1	20.2/20.6	17.5/18.0	32.5/32.5	18.2/17.2	21.3/21.3
M2	18.7/18.8	16.8/17.2	31.4/31.3	17.4/16.3	20.7/20.3
M3	18.0/18.1	16.0/16.7	30.8/30.7	17.0/15.9	20.0/19.8
M4	**17.7/17.6**	**15.0/16.0**	**30.5/30.4**	**16.2/15.0**	**18.9/18.7**

- **M0: Baseline**
- **M1: Baseline + SP**
- **M2: M1 + LP**
- **M3: M2 + Data Weighing**
- **M4: M3 + Dynamic Curriculum Learning**

- Three methods can be composed with others which achieve much better results

Results on Non Indo-European Languages

Methods	Tatar dev/test	Kabyle dev/test	Kinyarwanda dev/test
M0	26.6/27.1	53.4/53.1	48.3/48.5
M1	23.3/23.9	51.0/51.7	45.7/46.0
M2	18.5/18.7	43.0/42.9	42.6/42.7
M3	17.8/18.1	42.5/42.3	41.5/41.6
M4	**16.2/16.2**	**40.9/40.8**	**37.4/37.7**

Still Works!

- **M0 – M4 are the same as before**
- **Evaluate our methods on non Indo-European Languages**
- **Rich Resource languages are fr,pt,it,ca,eu(the same as basic setup)**

Thanks !