论文分享

ZiXi Yan

2022/06/10

TiBERT: Tibetan Pre-trained Language Model

- To better express the semantic information of Tibetan and reduce the problem of OOV, this paper uses the unigram language model of Sentencepiece to segment Tibetan words and constructs a vocabulary that can cover 99.95% of the words in the corpus.
- To further promote the development of various downstream tasks of Tibetan natural language processing, this paper collected a large-scale Tibetan dataset and trained the monolingual Tibetan pre-trained language model named TiBERT.
- To evaluate the performance of TiBERT, this paper conducts comparative experiments on the two downstream tasks of text classification and question generation. The experimental results show that the TiBERT is effective.

TiBERT: Tibetan Pre-trained Language Model

- Tibetan data from 21 Tibetan websites including Tibet Peo ple’s Network and Qinghai Provincial People’s Government Network

- The data contains knowledge in various fields such as current affairs, economy, technology, society, law, sports, life, nature, culture, geography, art, military, educa_x0002_tion, history, and people

TiBERT: Tibetan Pre-trained Language Model

TABLE IV
PERFORMANCES ON DOCUMENT CLASSIFICATION

Model	Accuracy(%)	Macro-Precision(%)	Macro-Recall(%)	Macro-F1(%)
CNN(syllable)	61.51	59.39	56.65	57.34
CINO-large	-	-	-	68.6
Transformer	28.63	41.21	28.63	28.79
TextCNN	61.71	61.65	61.71	61.53
DPCNN	62.91	63.61	62.91	61.17
TextRCNN	63.67	64.37	63.67	62.81
TiBERT	**71.04**	**71.20**	**71.04**	**70.94**
TiBERT+CNN	70.39	70.54	70.39	70.23

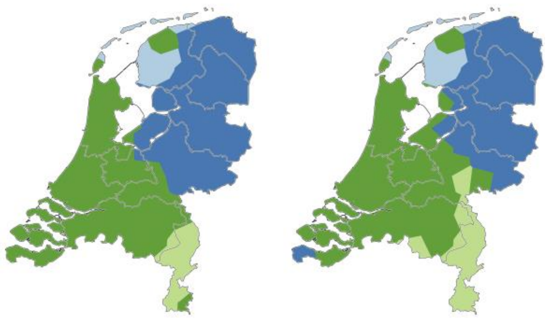
TiBERT: Tibetan Pre-trained Language Model

	Gold questions	S2S+ATT+CP	TiBERT
1	རང་བྱུང་ཁམས་ཀྱི་སྨུག་ཕྱིའི་འོད་ཀྱི་འབྱུང་ཁུངས་ནི་ཅི་ཞིག་ཡིན། What is the source of ultraviolet light in nature?)	རང་བྱུང་ཁམས་ཀྱི་སྨུག་ཕྱིའི་འོད་ཀྱི་འབྱུང་ཁུངས་ནི་གང་ཡིན། (What is the source of ultraviolet light in nature?)	རང་བྱུང་ཁམས་ཀྱི་སྨུག་ཕྱིའི་འོད་ཀྱི་འབྱུང་ཁུངས་ནི་གང་ཡིན། (What is the source of ultraviolet light in nature?)
2	མངའ་ཁོངས་ཀྱི་ཆ་སྙོམས་མཚོ་ངོས་མཐོ་ཚད་སྨི་ག་ཚོད་ཡིན། (What is the average elevation of the territory in meters?)	མངའ་ཁོངས་ཀྱི་ཆ་སྙོམས་**མཚོ་ངོས་མཚོ་ངོས་**མཐོ་ཚད་ག་ཚོད་ཡིན། (What is the average **elevation elevation** of the territory?)	མངའ་ཁོངས་ཀྱི་ཆ་སྙོམས་མཚོ་ངོས་མཐོ་ཚད་སྨི་ག་ཚོད་ཡོད། (What is the average elevation of the territory in meters?)
3	དགོན་གོང་མ་གྲོང་རྡལ་སྲིད་གཞུང་གི་ལྷོ་ནུབ་ཕྱོགས་སུ་ཏ་ལམ་སྤྱི་ལེ་ག་ཚོད་ཡོད། (How many kilometers southwest of Gongma town government?)	དགོན་གོང་མ་གྲོང་རྡལ་སྲིད་གཞུང་**གི་ལྷོ་ནུབ་ཕྱོགས་**གཞུང་**གི་ལྷོ་ནུབ་ཕྱོགས་**སུ་ཏ་ལམ་སྤྱི་ག་ཚོད་ཡོད། (How many centimeters is the **southwest southwest** of Gongma town government?)	དགོན་གོང་མ་གྲོང་རྡལ་སྲིད་གཞུང་གི་ལྷོ་ནུབ་ཕྱོགས་སུ་ཏ་ལམ་སྤྱི་ལེ་ཅི་ཙམ་མཚམས་ན་ཡོད། (How many kilometers southwest of Gongma town government?)
4	ངལ་རྩོལ་ནུས་ཤུགས་ཅན་གྱི་མི་ག་ཚོད་ཡོད། (How many people have the ability of work?)	**གྲོང་ཚོ་ཡོངས་ལ་སྡེ་ག་ཚོད་ཡོད།** (**How many villages are there in the whole town?**)	ངལ་རྩོལ་ནུས་ཤུགས་ཅན་གྱི་མི་ག་ཚོད་གི་**མི་ག་ཚོད་**ཡོད། (How **many many** people have the ability of work?)
5	དུས་ནམ་ཞིག་ལ་ཧུང་ཞེང་ཀུང་ཧྲེ་ལ་བསྒྱུར། (When will it be changed to Red Star Commune?)	དུས་ནམ་ཞིག་ལ་**ཀུང་ཧྲེ་ཀུང་ཧྲེ་ཀུང་ཧྲེ་**ལ་བསྒྱུར། (When will it be changed to a **commune commune commune**?)	སྤྱི་ལོ་**1966**ལོར་ཧུང་ཞེང་ཀུང་ཧྲེ་ལ་བསྒྱུར། (Changed to Red Star Commune in **1966**)

Quantifying Language Variation Acoustically with Few Resources

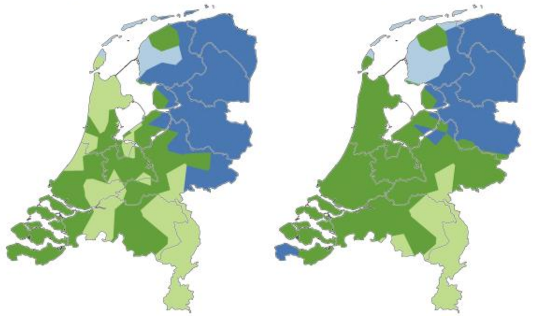
- 10 words (armen: ‘arms’ ,deeg: ‘dough’ , draden: ‘wires’ , duiven: ‘pigeons’ ,naalden: ‘needles’ , ogen: ‘eyes’ , pijpen: ‘pipes’ ,tangen: ‘pliers’ , volk: ‘people’ , vuur: ‘fire’)
- pro nounced in 106 locations in the Netherlands. On average, the duration of these 10 words is only 6.3 seconds for each location.

Quantifying Language Variation Acoustically with Few Resources



(a) Gold standard clustering (human-generated)

(b) XLSR-nl layer 15 (cl clustering)



(c) w2v2-en layer 13 (cl clustering)

(d) w2v2-nl layer 16 (wa clustering)

Figure 1: Cluster maps visualizing four clusters on the map of the Netherlands. Separate clusters are indicated by the different colours.

Automatic Pronunciation Assessment using Self-Supervised Speech Representation Learning

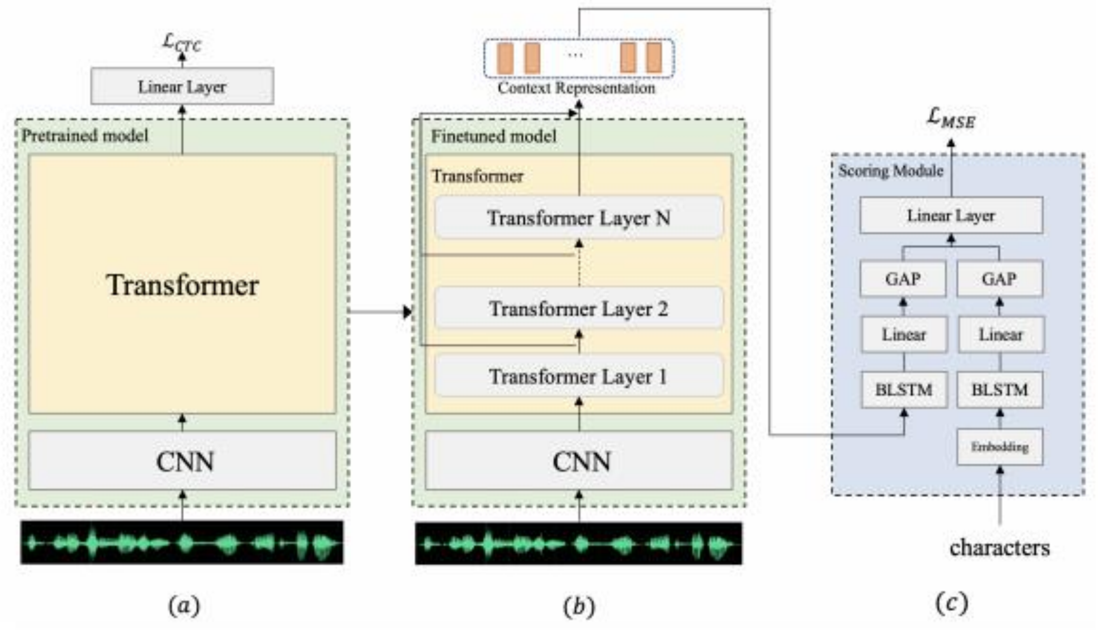


Figure 1: *Overall procedure of the proposed method for automatic pronunciation assessment based on SSL models.*

Automatic Pronunciation Assessment using Self-Supervised Speech Representation Learning

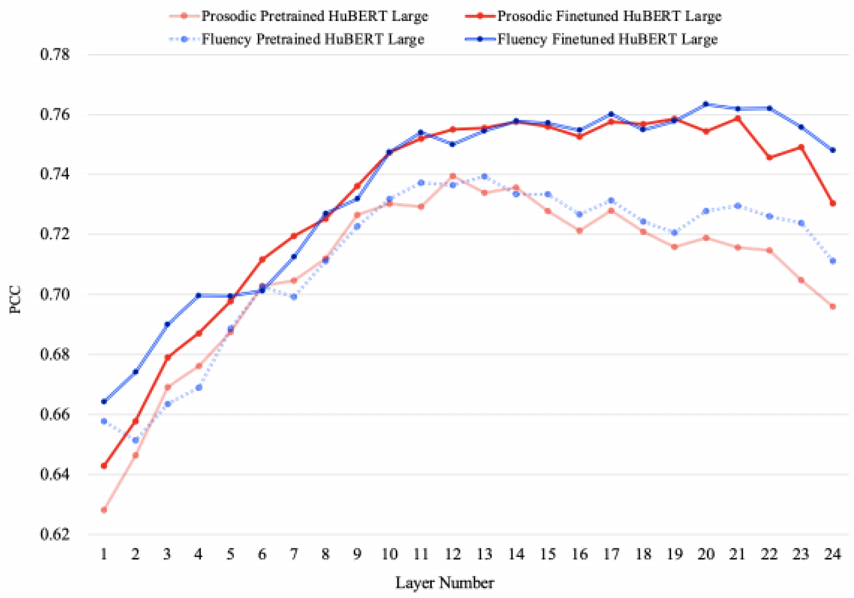


Figure 3: *PCCs when using the hidden states of different transformer layers of the fine-tuned HuBERT Large model as input to the BLSTM scoring module.*

Feature	KESL	Speechocean762	
	Holistic	**Fluency**	**Prosodic**
Local	0.56	0.60	0.62
Layer 20	0.81	0.76	0.76
All Layers (Proposed)	**0.82**	**0.78**	**0.77**

Table 2: *Comparison of the performance of local representation of a convolutional layer, the contextual representation, and layer-wise contextual representation of the transformer layers in HuBERT Large model.*

CATEGORICAL REPARAMETERIZATION WITH GUMBEL-SOFTMAX

The Gumbel-Max trick (Gumbel, 1954; Maddison et al., 2014) provides a simple and efficient way to draw samples z from a categorical distribution with class probabilities π:

$$z = \texttt{one\_hot}\left(\arg\max_i \left[g_i + \log \pi_i\right]\right) \tag{1}$$

where $g_1 \dots g_k$ are i.i.d samples drawn from $\text{Gumbel}(0, 1)$[1]. We use the softmax function as a continuous, differentiable approximation to $\arg\max$, and generate k-dimensional sample vectors $y \in \Delta^{k-1}$ where

$$y_i = \frac{\exp((\log(\pi_i) + g_i)/\tau)}{\sum_{j=1}^{k} \exp((\log(\pi_j) + g_j)/\tau)} \qquad \text{for } i = 1, \dots, k. \tag{2}$$

The density of the Gumbel-Softmax distribution (derived in Appendix B) is:

$$p_{\pi,\tau}(y_1, \dots, y_k) = \Gamma(k)\tau^{k-1}\left(\sum_{i=1}^{k} \pi_i / y_i^{\tau}\right)^{-k} \prod_{i=1}^{k} \left(\pi_i / y_i^{\tau+1}\right) \tag{3}$$

This distribution was independently discovered by Maddison et al. (2016), where it is referred to as the concrete distribution. As the softmax temperature τ approaches 0, samples from the Gumbel-Softmax distribution become one-hot and the Gumbel-Softmax distribution becomes identical to the categorical distribution $p(z)$.

CATEGORICAL REPARAMETERIZATION WITH GUMBEL-SOFTMAX

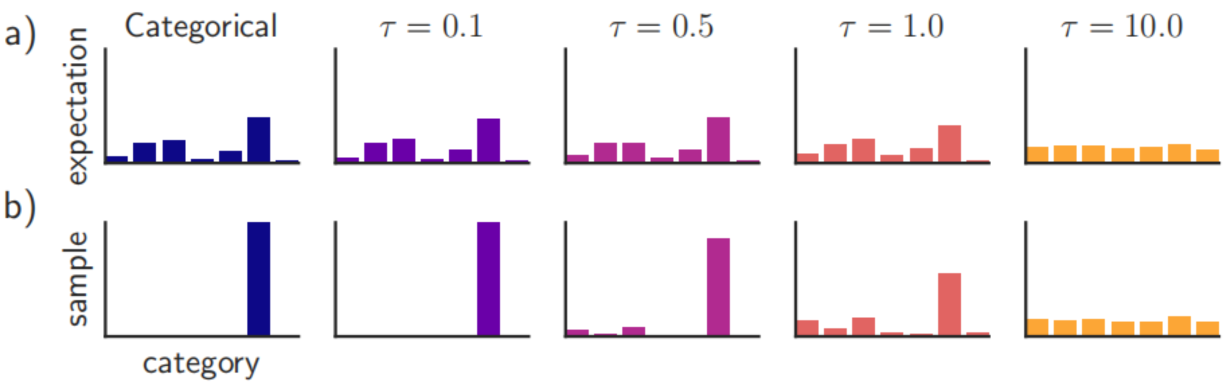


Figure 1: The Gumbel-Softmax distribution interpolates between discrete one-hot-encoded categorical distributions and continuous categorical densities. (a) For low temperatures ($\tau = 0.1, \tau = 0.5$), the expected value of a Gumbel-Softmax random variable approaches the expected value of a categorical random variable with the same logits. As the temperature increases ($\tau = 1.0, \tau = 10.0$), the expected value converges to a uniform distribution over the categories. (b) Samples from Gumbel-Softmax distributions are identical to samples from a categorical distribution as $\tau \to 0$. At higher temperatures, Gumbel-Softmax samples are no longer one-hot, and become uniform as $\tau \to \infty$.

CATEGORICAL REPARAMETERIZATION WITH GUMBEL-SOFTMAX

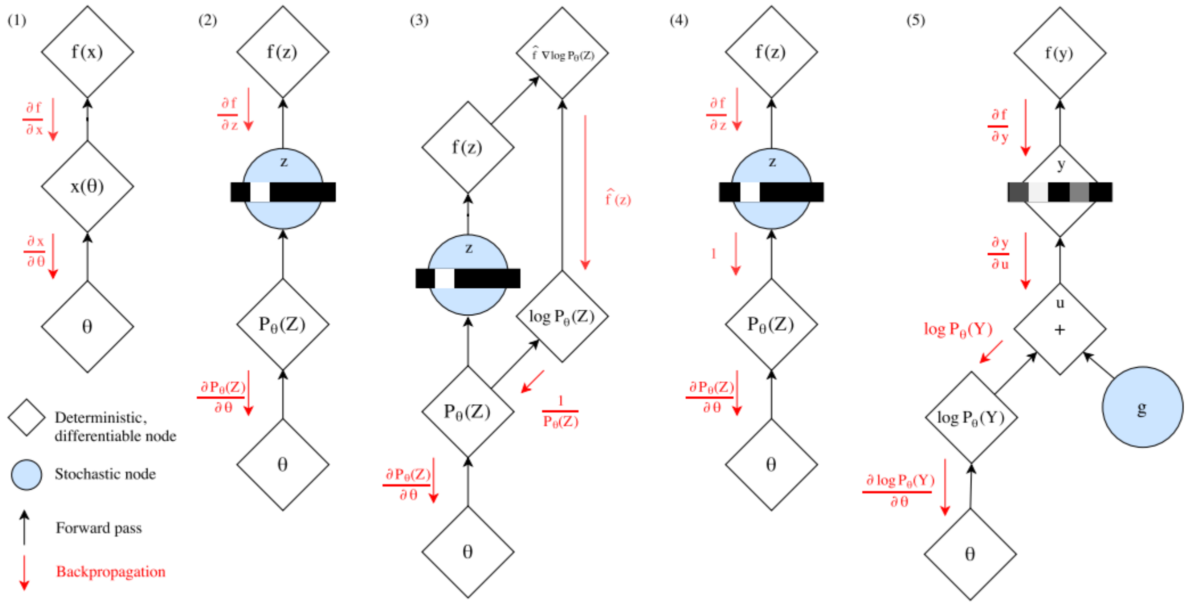


Figure 2: Gradient estimation in stochastic computation graphs. (1) $\nabla_\theta f(x)$ can be computed via backpropagation if $x(\theta)$ is deterministic and differentiable. (2) The presence of stochastic node z precludes backpropagation as the sampler function does not have a well-defined gradient. (3) The score function estimator and its variants (NVIL, DARN, MuProp, VIMCO) obtain an unbiased estimate of $\nabla_\theta f(x)$ by backpropagating along a surrogate loss $\hat{f} \log p_\theta(z)$, where $\hat{f} = f(x) - b$ and b is a baseline for variance reduction. (4) The Straight-Through estimator, developed primarily for Bernoulli variables, approximates $\nabla_\theta z \approx 1$. (5) Gumbel-Softmax is a path derivative estimator for a continuous distribution y that approximates z. Reparameterization allows gradients to flow from $f(y)$ to θ. y can be annealed to one-hot categorical variables over the course of training.

Thanks