IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 29, 2019, accepted May 15, 2019, date of publication May 27, 2019, date of current version June 17, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2919328

Morphological Verb-Aware Tibetan Language Model

KUNTHARRGYAL KHYSRU[1], DI JIN[1], (Member, IEEE), AND JIANWU DANG[1,2], (Member, IEEE)
[1]Tianjin Key Laboratory of Cognitive Computing and Application, College of Intelligence and Computing, Tianjin University, Tianjin 300350, China
[2]Japan Advanced Institute of Science and Technology, Ishikawa 9231292, Japan

Corresponding author: Di Jin (jindi@tju.edu.cn)

This work was supported in part by the National Key R & D Program of China under Grant 2018YFC0809800, and in part by the Natural Science Foundation of China under Grant 61772361.

ABSTRACT The Tibetan language model (TLM) is the key to Tibetan natural language processing. In this paper, we first observe that, different from widely used languages, Tibetan contains many morphological verbs that rarely appear in natural sentences but play a key role in accurate text prediction. This property is usually ignored by existing methods and makes traditional training strategies less effective in constructing accurate and robust TLMs. Hence, we propose a morphological verb-aware TLM by offline learning via a character frequency reweighting strategy and online tuning of discriminative weights conditioned on morphological verbs. However, because of the influence of morphological verbs on the tense and semantics of sentences, it is necessary to consider the morphological verbs in Tibetan. As a result, compared with state-of-the-art methods, our method not only reduces the perplexity but also improves the character error on tasks of the text prediction and automatic speech recognition (ASR).

INDEX TERMS Tibetan language model, text prediction, automatic speech recognition, morphological verb-aware model.

I. INTRODUCTION

Statistical language models (LMs) represent the probability that a sequence of words is a sentence and have been widely used in text prediction tasks, automatic speech recognition (ASR), machine translation, handwriting recognition and information retrieval [1]–[4]. N-gram LMs [5]–[7] are notable models that can be trained efficiently and have a powerful capacity to generalize. However, n-gram LMs usually struggle with modeling long-distance context dependencies. Recently, this problem has been significantly alleviated by modeling with a recurrent neural network (RNN)-based LM (RNNLM) [8]–[11].

However, an RNNLM is a state-of-the-art model that requires a large training dataset to learn effective parameters, leading to a severe data sparsity problem in which the language formation cannot be effectively represented by the LM due to the lack of training data or the key words rarely appearing;these key words are denoted as rare words. This situation is considerably severe for languages with low training resources. To alleviate this problem, recent works have attempted to reduce the number of trainable parameters by adding a compression layer into the LMs [7], [12] or performing data augmentation by adapting a large dataset to a small dataset. However, such methods cannot solve the problem introduced by rare words. Then, [12] and [13] proposed building an LM on the morpheme level via feature-rich modeling, which helps the RNN learn a more effective context for the LM. In addition, other structure information, e.g, subword, character, morph and morphology, have also been used to improve the LM [14]–[17], [19]. However, all of these works mainly focus on widely used languages, e.g., English or Chinese, while ignoring the specific properties of other low-resource languages, e.g., Tibetan. Hence, existing training methods for generic LMs still have great potential to be improved by carefully considering language-related properties.

There are few studies on the Tibetan language model (TLM), and the existing models basically use the n-gram method [30]. Recently, inspired by [8], [18] proposed building a TLM on the radical level instead of the character level to capture structural information from characters. Three types of embedded fusion methods have been proposed to enhance the model, including using uniform weight (TRU), different

The associate editor coordinating the review of this manuscript and approving it for publication was Mohammad Shorif Uddin.

weights (TRD) and radical combination (TRC). The purpose of our proposed model is to embed character embedding with a specific Tibetan radical unit and interpolation base and explore the different characteristics of radical embedding by introducing different radical embedding factors to make the model more flexible. Each radical embedding can be interpolated according to its contribution to the overall meaning of the corresponding character. In addition, there is a radical combination phenomenon in Tibetan. Radical embedding combination code can make full use of the nature of Tibetan radical embedding. Introduced by radical embedding, character embedding is more useful for enhancing semantic information, which can help solve data sparsity problems.

In this paper, we address the problem of constructing a TLM by focusing on Tibetan morphological verbs. Our contributions are three-fold. First, we study the importance of morphological verbs in TLM with the observation that: although morphological verbs rarely appear in Tibetan sentences, they constitute a large proportion of Tibetan words and play a key role in an effective TLM. Second, we propose an effective offline learning method by reweighting the character frequency, which results in a more powerful TLM. Third, we further enhance the importance of morphological verbs through online tuning of their discriminative weights to make an adaptive offline learning model. With the above efforts, compared with the baseline model, our new TLM achieves much better results on tasks of text prediction and ASR.

II. RELATED WORK

In this section, we investigate and discuss work related to LMs. We investigate work based on multiple granular methods and traditional methods and discuss the proposed state-of-the-art model for advancing the state-of-the-art.

A. DIFFERENT GRANULARITY OF INPUT

For low-resource languages, rare word processing is a common problem. [8] proposed a character embedding level-based model. They built a deeper network for reducing the traditional word embedding level to the character embedding level by the technology of a highway network, which avoided the problem of large-scale embedding computation and low-frequency words and obtained good performance. The model consisted of two parts, one that put the character embedding level as input to a convolutional neural network (CNN) and another that used the output of the CNN and the highway network as the input of an RNNLM. Furthermore, [4] also presented a model that combines an RNN with a character embedding level-based CNN.

In [8], by reducing the traditional word embedding level to the character level, large-scale embedding calculations and rare words were avoided, and a deeper network was constructed by highway network technology, which gave good results. The model consisted of two parts, the character level as input to the CNN, and input to the RNNLM through the output of the CNN and highway network, but the final prediction was still a word. Semantic and grammatical information can be obtained by experimenting with multiple language corpora as tests.

For morphologically rich languages, it is helpful to study the internal structural formation of words [31]–[33]. English morphology is the study of the relationship between the compositions of a word and the attempt to sort out the rules of its composition but without syntax and semantic information. Thus, language models can apply such information to obtain word sequences with high accuracy.

B. USING AN ADAPTIVE APPROACH

In an actual ASR task, a large number of domain matching datasets are not available. The LM requires domain-adaptive techniques to allow the use of multiple extradomain text resources [34], [35]. One of the most suitable methods for domain adaptation in language modeling is based on hybrid models [36], [37]. An adapted model can be constructed by combining LMs and hybrid weighting separately constructed from extradomain text resources [38]. In [39], because there was out-of-domain data, we applied larger data to adapt small data to solve the problem of an insufficient data volume.

To use the LM built from extradomain text resources for flexible domain adaptation, [28] and [29] developed a method for model merging in the potential variable space. In the latent variable space, a word is mapped to a potential variable space, so it can be expected to perform more flexible state sharing than possible in the observed word space. Thus, this paper introduces latent word LMs (LWLMs) into hybrid modeling [40]–[43]. The latent variables in the normal class of n-gram LMs are only model dependent indices, so each model has a different potential variable space [44], [45]. Therefore, the traditional class-based n-gram hybrid model must be performed in the observed text space [39], [46]. Additionally, latent variables in LWLMs are represented as specific potential words, and multiple LWLMs can share a common potential variable space, which enables us to perform flexible hybrid modeling while considering the potential variable space.

III. BACKGROUND

A. INTRODUCTION OF TIBETAN PROPERTIES

Tibetan belongs to the Tibeto-Burman language family, which is a language with a long history and great influence in the Sino-Tibetan language family. The Tibetan language in China is traditionally divided into the Lhasa dialect, Kham dialect and Amdo dialect. There is a certain difficulty in the difference between the Lhasa dialect and the Amdo dialect, and the Kham dialect is close to the Lhasa dialect. The internal differences in the Tibetan language are quite obvious. There are tones in the Lhasa and Kham dialects, but there are no tones in the Amdo dialect. Tibetan grammar is isolated, mainly relying on word order and auxiliary words to express various grammatical relations, and some verbs appear as inflections.

TABLE 1. Relationship between morphological verbs/morphologically invariant verbs and tense.

verb	future tense	present tense	past tense	Command tense	Word meaning
Morphological verbs	དཔྱད	ཚོད	དཔྱངས	ཚོད	"Finish"
	མཉན	ཉན	མཉན	ཉོན	"Listen"
	བསྒྱུར	སྒྱུར	བསྒྱུར	སྒྱུར	"Translation"
Morphologically invariant	བསྟོད	བསྟོད	བསྟོད	བསྟོད	"Praise"

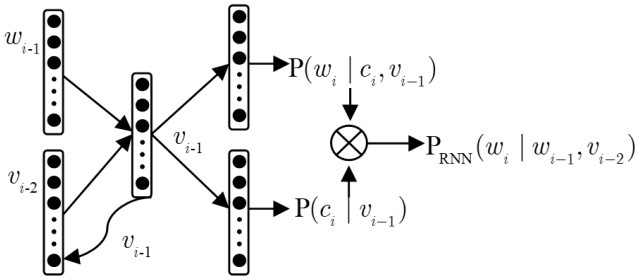


FIGURE 1. Structure of the baseline model: RNNLM.

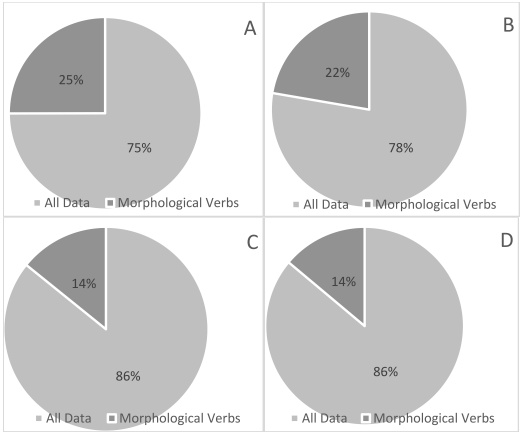


FIGURE 2. The distribution of the graph above in the corpus: A is the proportion of morphological verbs in the Tibetan audio corpus; B is the proportion of morphological verbs in the dictionary; C and D are the proportions of the training corpus and the testing morphological verbs.

Tibetan verbs are the key to understanding the meaning of a sentence and can be grouped into different subsets according to different properties. In particular, Tibetan verbs can be classified into morphological and morphologically invariant verbs according to their morphological properties. Specifically, morphological verbs may contain 2 to 4 tense variants, including future tense, present tense, past tense and command tense.

According to the characteristics of the changes in Tibetan verbs, we can see in Fig 1 and Fig 2 that they can be divided into morphological verbs and morphological nonvariable words. Morphological change verbs can be divided into four morphological changes, three morphological changes and two morphological changes. The four morphological changes refer to changes in character shape in the future, present time, completion time and command time. The three morphological changes mean that the morphological verb is generally one of a future time, present time, completion time and command time. The form is the same as one of the other three, but there is no law. The two form changes are the future, the present time, the completion time as well as the command time and one of the other three or two forms and the other. The two forms are the same, and there is no law.

Different from morphological verbs of English, morphological verbs of Tibetan are defined on the character level and contain additional tension, i.e, command tension, which is very important for sentence understanding [22]–[25]. Three cases of morphological verbs and one case of morphological invariant verbs are shown in Table 1. It can be seen that there are different changes in the sentences of different tense verbs, and these changes have an effect on the semantics of the sentence.

Due to the importance of morphological verbs for understanding Tibetan sentences, we propose a morphological verb-aware TLM for offline learning via character frequency reweighting and online tuning of the discriminative weight of morphological verbs. Experiments validate the effectiveness of our method.

B. BASELINE MODEL FOR TIBETAN

We can use an RNNLM [10] to the model Tibetan language. Specifically, to predict the current character w_i, we can encode the full history of the current character as $< w_{i-1,\ldots,}, w_1 >$, and compute the probabilities via a three layer RNN by

$$P_{RNN}(w_i| < w_{i-1,\ldots,}w_1 >) = P_{RNN}(w_i|w_{i-1}, v_{i-2}), \quad (1)$$

where v_{i-2} denotes the remaining historical context from 1 to $i-2$. We further modify Eq. (1) by adding a class-based factorized output layer structure. Each word in the output layer vocabulary is attributed to a unique class-based on a frequency count. We then obtain

$$\begin{aligned} P_{RNN}(w_i|w_{i-1}, v_{i-2}) &= P_{RNN}(w_i|v_{i-1}) \\ &= P(w_i|c_i, v_{i-1})p(c_i|v_{i-1}), \end{aligned} \quad (2)$$

c_i denotes the class label. $P(c_i|v_{i-1})$ is the conditional probability based class and is defined as a grouping result according to the character frequency on a training corpus. We also show the structure of Eq. (2) in Fig 2 and Fig 3 It has been demonstrated that an RNNLM with Eq. (2) can capture long-distance dependencies in English. $P(c_i|v_{i-1})$ plays a key role in such a model and relies on the classification of

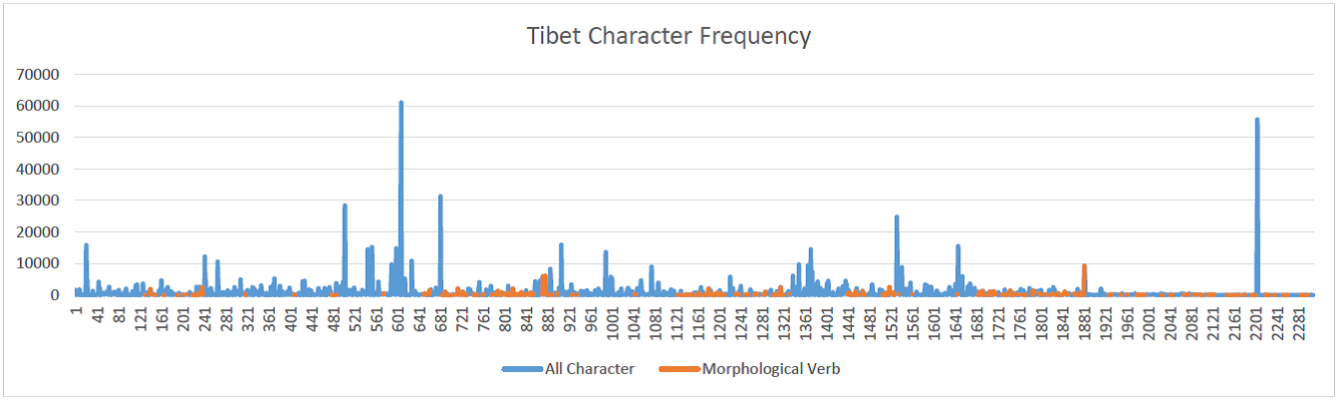


FIGURE 3. shows the frequency distribution of morphological verbs in our corpus.

characters based on a frequency count. This strategy does not consider the key to specific language properties, e.g, the importance of morphological verbs, in representing LMs.

IV. MORPHOLOGICAL VERB-AWARE TIBETAN LANGUAGE MODEL (TLM)

Verbs in natural language are indispensable parts of speech and are the core of sentence expression. The division of verbs is different because of language differences. Morphological verbs of Tibetan relate to not only semantics but also tense. Hence, it is very important to explore more effective training methods for TLMs considering morphological verbs with low resources.

A. KEY OF MORPHOLOGICAL VERBS IN TLM

The character is the smallest and most meaningful unit in Tibetan, analogous to words in English. As shown in Fig 2, we find that the proportions of morphological verbs in Tibetan are large, i.e., 25% and 22% in the corpus and dictionary (A and B in Fig 2), respectively, and 14% of the proportion of the training set and testing set are part of the morphological verbs (C and D in Fig 2). These verbs affect not only the temporal relationship of sentences but also the semantics of sentences.

However, as shown in Fig 3, the proportion of those important morphological verbs in our training corpus is low, which causes the trained TLM to easily miss semantic information from morphological verbs and directly affects the accuracy and speed of TLM-based recognition tasks. For example, we trained two TLMs based on a traditional RNNLM and our morphological verb-aware TLM on our corpus and used them to perform text prediction. Considering morphological verbs in both offline learning and online tuning, our method obtained a much lower prediction error.

In Table 2 and Fig 3, we can see that morphological verbs in Tibetan play an important role in the understanding of Tibetan sentences. In practice, because Tibetan is a low-resource language, audio and text data are limited. Therefore, we have limitations in obtaining morphological verb information in Tibetan. Therefore, it is necessary to increase the weight of morphological verbs to enhance such words and more accurately predict sentence semantics.

TABLE 2. Statistics of Tibetan text data.

Data set	#Token	% OOV
Char vocabulary	2472	-
STD	1.5m	1.08
LTD	21.3m	1.48
Valid set	125k	1.12
Test set	126k	1.11

B. OFFLINE LEARNING VIA CHARACTER FREQUENCY REWEIGHTING

In Eq. (2), $P(c_i|v_{i-1})$ is the conditional probability based on class c_i and is obtained according to the character frequency in the training corpus.c_i denotes the class of character w_i in the training corpus.$P(c_i|v_{i-1})$ provides a discriminative weight, i.e, a prior, about the prediction of w_i given historical information v_{i-1} and helps the RNN model estimate more accurately. However, such a prior based character frequency ignores the importance of morphological verbs in understanding sentences. To overcome this problem, we propose reweighting the character frequency of morphological verbs in the training corpus.

Specifically, after we calculate the character frequency of all characters in the corpus, we reweight the character frequency of morphological verbs by multiplying their values by a weight β. In this paper, we set $\beta = 3$.We then classify all of the characters according to this new reweighted character frequency, thus leading to a new $P(c_i|v_{i-1})$. Thus, some of the original low-frequency morphological verbs can be divided into new higher classes to improve the prediction speed and accuracy. This method is called RNNLM_character frequency reweighting (_CFR).

C. ONLINE TUNING DISCRIMINATIVE WEIGHTS

To further understand morphological verbs during testing, we propose online tuning of the discriminative weights, i.e,$P(c_i|v_{i-1})$, to improve the prediction accuracy. The entire

TABLE 3. The basic information on the Tibetan audio data.

Number of Speakers		Speech signal		Number of text	
Male	Female	Sampling Rate	Quantification precision	Train set	Test set
13	10	16K HZ	16-bit	36,090	2,644

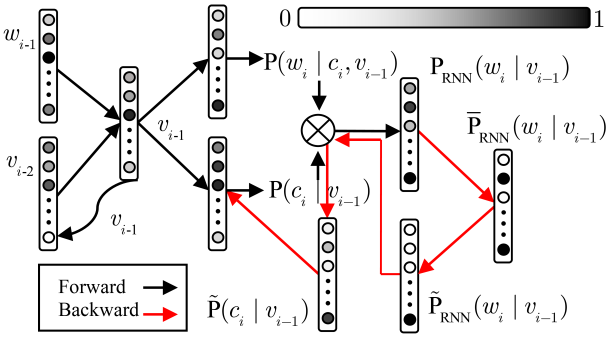


FIGURE 4. An example of online tuning discriminative weights.

process is shown in Fig 4. Specifically, the output of an RNNLM is given to a threshold ε. Then, a binary output is generated

$$\bar{P}_{RNN}(w_i|v_{i-1}) = P_{RNN}(w_i|v_{i-1}) > \varepsilon. \quad (3)$$

The values of $\bar{P}_{RNN}(w_i|v_{i-1})$ that do not belong to the morphological verb are set to 0, and $\tilde{P}_{RNN}(w_i|v_{i-1})$. We regard $\tilde{P}_{RNN}(w_i|v_{i-1})$ as the prediction of the RNNLM and back-project it to $P(c_i|v_{i-1})$ and obtain

$$\tilde{P}(c_i|v_{i-1}) = \frac{\tilde{P}_{RNN}(w_i|v_{i-1})}{P(w_i|c_i, v_{i-1})}. \quad (4)$$

We generate a new $P(c_i|v_{i-1})$ by combining it with $\tilde{P}(c_i|v_{i-1})$

$$\bar{P}(c_i|v_{i-1}) = P(c_i|v_{i-1}) + \alpha\tilde{P}(c_i|v_{i-1}), \quad (5)$$

where α denotes the combination weight of $\tilde{P}(c_i|v_{i-1})$. We denote this method as RNNLM_tuning discriminative weights (_TDW).

V. EXPERIMENTS

In the experimental section, we first introduce the experimental corpus, including the speech corpus and the text corpus of the Lhasa dialect. Then, based on the baseline LM modeling method and our proposed method, we further compare the existing methods with our results in different hidden layers.Next, we interpolate our method with the traditional n-gram method and compare it with the results of other interpolations. In the experiment, we denoted the Kneser-Ney smoothed 3-gram as KN3. Finally, applying our method to ASR verifies the effectiveness of our approach.

A. SETUP

Through experiments, we verify the performance of our method in the Tibetan language. We address the character

TABLE 4. Comparison between the PPL of the previous method and our method compare on the same topic.

Granularity	Language Model	PPL
character	N-gram(Mikolov et al.2012)	55.2
	RNNLM(Mikolov et al.2012)	62.9
	CUED_RNNLM(Xie Chen et al.2016)	58.4
	LSTM(Xie Chen et al.2016)	55.9
	CharCNN (Kim et al., 2016)	55.2
radical	_TRU(Shen et al.2017)	57.6
	_TRD(Shen et al.2017)	54.3
	_TRC(Shen et al.2017)	53.8
morph	**_CFR**	**50.6**
	_TDW	**49.8**

TABLE 5. Comparison of the PPL on the same topic after interpolation with our method.

Granularity	Language Model + KN3	PPL
character	N-gram(Mikolov et al.2012)+ KN3	55.2
	RNNLM(Mikolov et al.2012)+KN3	48.2
	CUED_RNNLM(Xie Chen et al.2016)+KN3	48.0
	LSTM(Xie Chen et al.2016)+KN3	46.4
	CharCNN (Kim et al.,2016)+KN3	47.3
radical	_TRU(Shen et al.2017)+KN3	47.9
	_TRD(Shen et al.2017)+KN3	47.0
	_TRC(Shen et al.2017)+KN3	46.9
morph	**_CFR+KN3**	**45.1**
	_TDW+KN3	**44.2**

as input. In the experiment, we chose perplexity (PPL) and character error rates (CERs) as the evaluation criteria.

The choice of corpus directly affects the quality of the TLM, which affects the speech recognition performance. For LMs, sentences that are commonly used in real life should be chosen to conform to the habits of people using natural language. There is no open database for the Tibetan language. In this paper, we evaluated news data from the Internet to build an LM, and then divided it into a training set, a validation set and a testing set in a 10:1:1 ratio [8], [17]. This small Tibetan Training dataset (STD), is in-domain with the testing set. In a large Tibetan training dataset (LTD), a word is limited to the first 2,472 characters based on the frequency of the audio data. In addition, out-of-vocabulary(OOV) symbols are used to render any character that is not in part of the selected vocabulary. The size of the corpus and the percentage of OOV characters are shown in Table 2.

As a minority language in China, the Tibetan language has a limited scope of application. Coupled with the government's policy reasons, more data are biased towards the news because of the limited corpus and the scattered theme. In the experiment we have two data sets, STD and LTD, but in this

TABLE 6. Our method on the STD data set and interpolation on the N-gram interpolation and LTD data sets.

# hidden	RNNLM+KN3(S)	RNNLM+KN3(L)	_TRU+KN3(S)	_TRU+KN3(L)	_CFR+KN3(S)	_TDW +KN3(S)	_CFR+KN3(L)	_TDW +KN3(L)
400	47.9	46.5	48.8	47.4	**46.2**	**45.7**	**44.6**	**43.8**
500	47.2	45.5	48.4	46.7	**45.9**	**44.9**	**44.2**	**42.9**
600	48.5	46.9	48.3	46.7	**45.7**	**44.5**	**44.1**	**42.3**
700	48.7	46.8	48.0	45.9	**45.1**	**44.2**	**43.2**	**42.1**

experiment we focus on the application of STD data, and the LTD is the auxiliary. The STD data are biased towards news, while the LTD data topics are scattered, including: news, politics, economics, culture, Buddhism and Gesar.

Our speech corpus is derived from speakers who are college students whose mother tongue is the Lhasa dialect. It was collected from 23 Tibetan native speakers, including 13 males and 10 females. All speech signals were sampled at 16 KHz with 16-bit quantization. For the purpose of building a practical ASR system, the recording scripts consist of mainly declarative sentences covering wide topics.

There are more than 38,700 sentences in the corpus; among them, 36,090 sentences are used in the training set, and 2,664 sentences are used in the testing set. There is no overlap between the training and testing sets in utterances and speakers. Table 3 shows the basics of our audio corpus.

B. COMPARISON RESULTS ON TEXT PREDICTION

Table 4 shows the result of our latest method on the STD dataset. Based on an RNNLM, we use the radical-based Tibetan radical uniform weight (_TRU) method as the baseline. Our method is approximately 15.5% less than the RNNLM and 11.6% less than the baseline _TRU method, and confusion is reduced compared to the RNNLM Tibetan radical different weight (_TRD) method and the Tibetan radical combination weight (_TRC) method by approximately 6.8% and 5.8%, respectively, indicating the effectiveness of our method [18].Our method's confusion is reduced by 8.3% compared to the traditional n-gram method.

The RNNLMs and n-gram LMs have their modeling characteristics as two essentially different LMs. RNNLMs typically use a fixed weight of linear interpolation in conjunction with the n-gram LM. Table 5 shows the result of our method and n-gram interpolation on the STD dataset, referring to the value of λ in [17], [22], [23] ($\lambda = 0.5$).

In Table 5, we can see that the proposed method achieves better results than the traditional n-gram method. In addition, when our method is combined with the n-gram method, some improvements are achieved. This proves that our method and the n-gram method have complementary contributions to solve the problem of rate words, which further proves the effectiveness of our improved method in solving the sparseness problem of Tibetan data.

Table 6 is the result of our interpolation of STD data and LTD data. KN3(S) refers to the application of the n-gram model to small data training, while KN3(L) refers to the model utilized for larger data training. For STD data, our method and n-gram are reduced by 7.4% and 9.3%, respectively, compared to the RNNLM method and n-gram. Our method and n-gram are 6.1% less than the _TRU method, and n-gram confusion is lower by approximately 7.9%. The n-gram model trained on the LTD data is different from our method, which are 11.3% and 13.6% less than the RNNLM method and the n-gram method, respectively, which are better than the _TRU and the n-gram methods. Confusion is reduced by 10% and 12.3%.

TABLE 7. The latest method and the %CER of our method.

Granularity	Language Model	%CER
character	N-gram(Mikolov et al.2012)	35.20
	RNNLM(Mikolov et al.2012)	34.60
	CUED_RNNLM(Xie Chen et al.2016)	34.25
	LSTM(Xie Chen et al.2016)	33.96
	CharCNN (Kim et al.,2016)	34.03
radical	_TRU(Shen et al.2017)	34.09
	_TRD(Shen et al.2017)	34.15
	_TRC(Shen et al.2017)	33.94
morph	**_CFR**	**33.55**
	_TDW	**33.10**

C. COMPARISON RESULTS ON SPEECH RECOGNITION

It can be seen that the proposed method of _CFR and N-gram(S) difference results in better results than the differences of other methods. In particular, our proposed difference method between _TDW and N-gram(L) not only achieves better results than the existing method, but also increases the relative _CFR method by approximately 4.5% in PPL. This result also shows that our method has a good effect on the improvement of the Tibetan language model.

The above experiments are based on PPL as the evaluation criterion to verify the results. We can see that our method, rather than the latest TLM research, achieved better results. To verify the validity of our proposed method, we used the experimental results of different granularity and our proposed method for application to ASR for verification. The results show that the RNNLM method is better than the traditional n-gram method regarding character granularity.

The _TRC method for Tibetan in radical granularity achieved the best results. The results in Table 7 show a _CFR-based RNNLM relative improvement of %CER from 3.1%. Using the _TDW-based RNNLM relative improvement, the %CER improves from 4.3%. The method we

TABLE 8. Evaluation result of the %CER with N-best rescoring.

N	# hidden units	%CER with N-best rescoring						Original
		RRNNLM	_TRU	_TRD	_TRC	_CFR	_TDW	
100	500	33.84	34.22	34.07	34.02	**33.65**	**33.20**	35.20
	600	34.03	33.99	34.06	34.05	**33.55**	**33.10**	
	700	33.97	34.09	34.15	33.94	**33.55**	**33.03**	
1000	500	33.73	34.15	33.92	34.02	**32.87**	**32.71**	
	600	33.88	33.87	34.06	33.82	**32.87**	**32.65**	
	700	33.83	34.05	34.08	33.78	**32.74**	**32.55**	

TABLE 9. Subjective evaluation after a comparison between the baseline model and our model.

Number	1	2	3	4	5
Semantic influential number of sentences	46	49	45	48	47

propose has a good effect on Tibetan speech recognition compared with the latest method.

We know that a lattice is a structure decoded once in the speech recognition process that contains a large number of candidate results. Since the neural network uses historical information to predict the next word, rerating the lattice will result in slow search speeds. Compared to the word structure of the lattice, N-best is more suitable for the model extension of long-distance information. This paper uses the intermediate result of N-best for rescoring [24], as shown in Table 8.

We validated our model on the ASR experiment in the Tibetan Lhasa dialect audio dataset [20]. Table 8 is the result of our verification of %CER in Tibetan. Our method is reduced by approximately 3.5% in the N-best (n=100 and n=1,000) rescores, indicating that our method has a good effect on the weighting method of rare words in the TLM.

D. DISCUSSION

The above experimental results show that the proposed RNNLM_character frequency reweighting (_CFR) and the RNNLM_tuning discriminative weights (_CFR) are effective. Some interesting observations are made as follows.

All the research on the LM rate word is not the same; some scholars have developed weighted methods based on rate words, and some scholars have proposed adaptive methods. The question with these studies is whether all of these rate words make sense. Therefore, according to the characteristics of Tibetan morphological verbs, we propose a morphological verb-aware TLM. The use of morphological verb weighting can not only affect the change in type but also improve the predictive capacity. Therefore, increasing the weight of morphological verbs can learn more language features. The language features are helpful for improving the performance of the TLM.

In Tibetan, the morphological verb has a greater influence on the tense of the sentence, and there is no uniform standard in temporal changes. In Table 1, the morphological verbs in Tibetan can be divided into morphological verbs and morphologically invariant verbs. The former is the focus of our research, because such verbs play a major role in the change of the tense in the sentence, while the latter is the same as the characters we generally encounter. However, these characters appear less frequently in the training corpus, which affects the predictive power of the sentence, especially in speech recognition tasks. Therefore, we need to conduct research and analysis on such characters to find a more accurate way to solve such problems.

To validate our approach, we experimented on ASR. We tested 100 sentences in the set baseline output semantics that are inaccurate, output them in the way we proposed, and then compared them with the original sentences. The method of comparison was to use subjective evaluation, and we looked for 5 experts who worked in the Tibetan language to score. Table 9 shows the results of the comparison between the model and the baseline model; the criterion of the score was the original sentence as the standard, and the semantics of the 100 sentences were the most similar to the original sentence. As shown in Table 9, our proposed method was better than the baseline method in semantic understanding, and in the 100 sentences output in the baseline model, our model could accurately represent 47% of the sentence semantics.

In summary, the weighted method based on morphological verbs influences the semantics of sentences to a certain extent and achieves good results. Therefore, it is necessary to strengthen morphological verbs in Tibetan.

VI. CONCLUSION AND FERYURE WORK

In this paper, we have shown that Tibetan morphological verbs are rare words that are very important for learning an effective TLM, and we have proposed a morphological verb-aware TLM. We first proposed an offline learning TLM by character frequency reweighting to enhance the weights of morphological verbs. Furthermore, we proposed online tuning of the discriminative weights of morphological verbs to make the offline learned TLM online adaptive. As a result, our method outperforms baseline models on tasks of text prediction and ASR.

ACKNOWLEDGMENT

We would also like to thank Qing Guo for his contribution for this research.

REFERENCES

[1] P. F. Brown, J. Cocke, S. A. D. Pietra, V. J. D. Pietra, F. Jelinek, J. D. Lafferty, and P. S. Roossin, "A statistical approach to machine translation," *Comput. Linguistics*, vol. 16, no. 2, pp. 79–85, 1990.

[2] C. Zhai and J. Lafferty, "A study of smoothing methods for language models applied to information retrieval," *ACM Trans. Inf. Syst.*, vol. 22, no. 2, pp. 179–214, Apr. 2004.

[3] R. Kuhn and R. de Mori, "A cache-based natural language model for speech recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, no. 6, pp. 570–583, Jun. 1990.

[4] R. Jozefowicz, O. Vinyals, M. Schuster, N. Shazeer, and Y. Wu, "Exploring the limits of language modeling," Feb. 2016, *arXiv:1602.02410*. [Online]. Available: https://arxiv.org/abs/1602.02410

[5] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," *Comput. Speech Lang.*, vol. 13, no. 4, pp. 359–394, Oct. 1999.

[6] B. Roark, M. Saraclar, and M. Collins, "Discriminative n-gram language modeling," *Comput. Speech Lang.*, vol. 21, no. 2, pp. 373–392, 2007.

[7] P. F. Brown, P. V. DeSouza, R. L. Mercer, V. J. D. Pietra, and J. C. Lai, "Class-based n-Gram models of natural language," *J. Comput. Linguistics*, vol. 18, no. 4, pp. 467–479, 1992.

[8] Y. Kim, Y. Jernite, D. Sontag, and A. M. Rush, "Character-aware neural language models," in *Proc. AAAI*, 2016, pp. 2741–2749.

[9] T. Mikolov, M. Karafiát, L. Burget, J. Černocký, and S. Khudanpur, "Recurrent neural network based language model," in *Proc. 11th Annu. Conf. Int. Speech Commun. Assoc.*, Chiba, Japan, 2010, pp. 1045–1048.

[10] T. Mikolov, S. Kombrink, L. Burget, J. Černocký, and S. Khudanpur, "Extensions of recurrent neural network language model," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2011, pp. 5528–5531.

[11] M. Sundermeyer, H. Ney, and R. Schlüter, "From feedforward to recurrent LSTM neural networks for language modeling," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 3, pp. 517–529, Mar. 2015.

[12] A. E.-D. Mousa, H.-K. J. Kuo, L. Mangu, and H. Soltau, "Morpheme-based feature-rich language models using deep neural networks for LVCSR of Egyptian Arabic," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)* Vancouver, BC, Canada, May 2013, pp. 8435–8439.

[13] Y. Shi, P. Wiggers, and C. M. Jonker, "Towards recurrent neural networks language models with linguistic and contextual features," in *Proc. 13th Annu. Conf. Int. Speech Commun. Assoc.*, 2012, pp. 1662–1665.

[14] A. E.-D. Mousa, R. Schlüter, and H. Ney, "Investigations on the use of morpheme level features in language models for arabic LVCSR," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2012, pp. 5021–5024.

[15] Y. Z. He, B. Hutchinson, P. Baumann, M. Ostendorf, E. Fosler-Lussier, and J. Pierrehumbert, "Subword-based modeling for handling OOV words inkeyword spotting," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Florence, Italy, May 2014, pp. 7864–7868.

[16] T. He, X. Xiang, Y. Qian, and K. Yu, "Recurrent neural network language model with structured word embeddings for speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 5396–5400.

[17] X. Chen, X. Liu, A. Ragni, Y. Wang, and M. J. F. Gales, "Future word contexts in neural network language models," in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop (ASRU)*, Dec. 2017, pp. 97–103.

[18] T. Shen, L. Wang, X. Chen, K. Khysru, and J. Dang, "Exploiting the tibetan radicals in recurrent neural network for low-resource language models," in *Proc. Int. Conf. Neural Inf. Process.*, 2017, pp. 266–275.

[19] H. Fang, M. Ostendorf, P. Baumann, and J. Pierrehumbert, "Exponential language modeling using morphological features and multi-task learning," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 12, pp. 2410–2421, Dec. 2015.

[20] T. Shen, L. Wang, X. Chen, K. Khuru, and J. W. Dang, "Tibetan language model based on recurrent neural network," *Int. Seminar Speech Prod.,* 2017.

[21] T. Shen, L. Wang, X. Chen, K. Khuru, and J. Dang, "Investigation of long short-term memory for tibetan language model," in *Proc. Nat. Conf. Man-Mach. Speech Commun.*, 2017.

[22] X. Chen, X. Wang, X. Liu, M. J. F. Gales, and P. C. Woodland, "Efficient GPU-based training of recurrent neural network language models using spliced sentence bunch," in *Proc. 15th Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2014, pp. 641–645.

[23] X. Chen, X. Liu, Y. Qian, M. J. F. Gales, and P. C. Woodland, "CUED-RNNLM—An open-source toolkit for efficient training and evaluation of recurrent neural network language models," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Shanghai, China, Mar. 2016, pp. 6000–6004.

[24] X. Liu, X. Chen, Y. Wang, M. J. Gales, and P. C. Woodland, "Two efficient lattice rescoring methods using recurrent neural network language models," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 8, pp. 1438–1449, Aug. 2016.

[25] *Luosang Tsechum Gyumsto Seduo, Tibetan Grammatical Theories by Seduo*. 1st ed., Nationalities Publishing house, Feb. 1957.

[26] X. Tsedan, *Detailed Explanation About Tibetan Grammar*, 1st ed. Qinghai Sheng, China: Qinghai Nationalities Publishing House, 1954.

[27] *Lcagsuntharrgyal, A Study of Tibetan Grammar*, Qinghai Nationalities Publishing House, Qinghai Sheng, China, 2008.

[28] R. Masumura, T. Asami, T. Oba, H. Masataki, S. Sakauchi, and A. Ito, "Domain adaptation based on mixture of latent words language models for automatic speech recognition," *IEICE Trans. Inf. Syst.*, vol. 101, no. 6, pp. 1581–1590, 2018.

[29] R. Masumura, T. Asami, T. Oba, H. Masataki, and S. Sakauchi, "Viterbi approximation of latent words language models for automatic speech recognition," *J. Inf. Process.*, vol. 27, pp. 168–176, 2019.

[30] J. Li, H. Wang, L. Wang, J. Dang, K. Khuru, and G. Lobsang, "Exploring tonal information for Lhasa dialect acoustic modeling," in *Proc. 10th Int. Symp. Chin. Spoken Lang. Process. (ISCSLP)*, Oct. 2016, pp. 1–5.

[31] A. Lazaridou, M. Marelli, R. Zamparelli, and M. Baroni, "Compositionally derived representations of morphologically complex words in distributional semantics," in *Proc. 51st Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2013, pp. 1517–1526.

[32] T. Luong, R. Socher, and C. Manning, "Better word representations with recursive neural networks for morphology," in *Proc. 17th Conf. Comput. Natural Lang. Learn.*, 2013, pp. 104–113.

[33] E. Yildiz, C. Tirkaz, H. B. Sahin, M. T. Eren, and O. O. Sonmez, "A morphology-aware network for morphological disambiguation," in *Proc. 13th AAAI Conf. Artif. Intell.*, 2016, pp. 1–7.

[34] J. R. Bellegarda, "Statistical language model adaptation: Review and perspectives," *Speech Commun.*, vol. 42, pp. 93–108, Jan. 2004.

[35] P. Koehn and J. Schroeder, "Experiments in domain adaptation for statistical machine translation," in *Proc. 2nd Workshop Stat. Mach. Transl.*, 2007, pp. 224–227.

[36] S. Katz, "Estimation of probabilities from sparse data for the language model component of a speech recognizer," *IEEE Trans. Audio, Speech Lang. Process.*, vol. ASP-35, no. 3, pp. 400–401, Mar. 1987.

[37] R. M. Iyer and M. Ostendorf, "Modeling long distance dependence in language: Topic mixtures versus dynamic cache models," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 1, pp. 30–39, Jan. 1999.

[38] R. Iyer, M. Ostendorf, and H. Gish, "Using out-of-domain data to improve in-domain language models," *IEEE Signal Process. Lett.*, vol. 4, no. 8, pp. 221–223, Aug. 1997.

[39] T. R. Niesler and P. C. Woodland, "Combination of word-based and category-based language models," in *Proc. ICSLP*, vol. 1, Oct. 1996, pp. 220–223.

[40] K. Deschacht, J. De Belder, and M.-F. Moens, "The latent words language model," *Comput. Speech Lang.*, vol. 26, no. 5, pp. 384–409, 2012.

[41] R. Masumura, H. Masataki, T. Oba, O. Yoshioka, and S. Takahashi, "Use of latent words language models in ASR: A sampling-based implementation," in *Proc. ICASSP*, May 2013, pp. 8445–8449.

[42] R. Masumura, T. Oba, H. Masataki, O. Yoshioka, and S. Takahashi, "Viterbi decoding for latent words language models using Gibbs sampling," in *Proc. INTERSPEECH*, 2013, pp. 3429–3433.

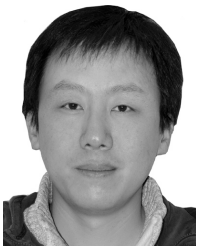
[43] R. Masumura, T. Adami, T. Oba, H. Masataki, S. Sakauchi, and S. Takahashi, "N-gram approximation of latent words language models for domain robust automatic speech recognition," *IEICE Trans. Inf. Syst.*, vol. E99-D, no. 10, pp. 2462–2470, 2016.

[44] S. Goldwater and T. Griffiths, "A fully Bayesian approach to unsupervised part-of-speech tagging," in *Proc. ACL*, 2007, pp. 744–751.

[45] P. Blunsom and T. Cohn, "A hierarchical Pitman-Yor process HMM for unsupervised part of speech induction," in *Proc. ACL*, 1996, pp. 865–874.

[46] R. C. Moore and W. Lewis, "Intelligent selection of language model training data," in *Proc. ACL*, 2010, pp. 220–224.

KUNTHARRGYAL KHYSRU received the bachelor's degree in art design and the master's degree in Tibetan information processing from Qinghai Nationalities University, Qinghai, China, in 2005 and 2012, respectively. His research interests include Tibetan signal processing and Tibetan natural language processing.



DI JIN received the B.S., M.S., and Ph.D. degrees from Jilin University, Changchun, China, in 2005, 2008, and 2012, respectively, all in computer science. He was a Postdoctoral Research Fellow with the School of Design, Engineering, and Computing, Bournemouth University, Poole, U.K., from 2013 to 2014. He is currently an Associate Professor with the College of Intelligence and Computing, Tianjin University, Tianjin, China. He has published over 40 international journal articles and conference papers. His current research interests include data mining, machine learning, and natural language processing.

JIANWU DANG received the B.E. and M.E. degrees from Tsinghua University, China, in 1982 and 1984, respectively, and the Ph.D. degree from Shizuoka University, Japan, in 1992. He was with Tianjin University, Tianjin, China, as a Lecturer, from 1984 to 1988. From 1992 to 2001, he was with ATR Human Information Processing Labs., Japan. Since 2001, he has been on the Faculty of the School of Information Science, Japan Advanced Institute of Science and Technology (JAIST), as a Professor. He joined the Institute of Communication Parlee (ICP), Center of National Research Scientific, France, as a Research Scientist, from 2002 to 2003. Since 2009, he has been with Tianjin University.