Effective Training End-to-End ASR systems for Low-resource Lhasa Dialect of Tibetan Language

Lixin Pan*, Sheng Li†, Longbiao Wang* and Jianwu Dang‡
* Tianjin University, Tianjin, China
E-mail: panlixin, longbiao_wang@tju.edu.cn
† National Institute of Information and Communications Technology, Kyoto, Japan
E-mail: sheng.li@nict.go.jp
‡ Japan Advanced Institute of Science and Technology, Ishikawa, Japan
E-mail: jdang@jaist.ac.jp

Abstract—**The Lhasa dialect is the most important Tibetan dialect and has the largest number of speakers in Tibet and massive written scripts in the long history. Studying how to apply speech recognition techniques to Lhasa dialect has special meaning for preserving Tibet's unique linguistic diversity. Previous research on Tibetan speech recognition focussed on selecting phone-level acoustic modeling units and incorporating tonal information but paid less attention to the problem of limited data. In this paper, we focus on training End-to-End ASR systems for Lhasa dialect using transformer-based models. To solve the low-resource data problem, we investigate effective initialization strategies and introduce highly compressed and reliable sub-character units for acoustic modeling which have never been used before. We jointly training the transformer-based End-to-End acoustic model with two different acoustic unit sets and introduce an error-correction dictionary to further improve the system performance. Experiments show our proposed method can effectively modeling low-resource Lhasa dialect compared to DNN-HMM baseline systems.**

I. Introduction

Tibet's culture is undergoing drastic modernization transformations in the twenty-first century. How to preserve Tibet's unique linguistic diversity is a very challenging topic today. Among the numerous spoken forms of the Tibetic language family, there are three major dialects: Lhasa Tibetan, Khams Tibetan, and Amdo Tibetan. The Lhasa (the central Tibetan dialect) is the most influential dialect and has the largest number of speakers. Most of the classic Tibetan manuscripts were written with this language in the long history. For this reason, studying how to apply natural language processing and speech recognition techniques to Lhasa dialect has drawn increasing attention.

Conventional automatic speech recognition (ASR) systems (GMM-HMM [1] and DNN-HMM [2]) require independently optimized components: acoustic model, lexicon and language model. The previous works on Tibetan speech recognition research focussed on selecting acoustic modeling units [3], incorporating effective tonal information [4], using the lattice-free maximum mutual information (LFMMI) [5] and transfer-learning [6] to enhance Tibetan ASR systems. However, the improvement is still limited due to the low-resource data problem. It is necessary to study novel acoustic modeling techniques for Tibetan to further promote the performance of Tibetan speech recognition.

The End-to-End neural network model simplified the ASR system construction, and solved the sequence labeling problem between variable-length speech frame inputs and label outputs (phone, character, syllable, word, etc.) and achieved promising results on ASR tasks. Various types of End-to-End model have been studied in recent years, i.e. connectionist temporal classification (CTC) [7], [8], attention-based encoder-decoder (Attention) End-to-End models [9], [10], End-to-End LFMMI [11] and End-to-End models jointly trained with CTC and Attention objectives (CTC/Attention) [12], [13], [14], [15]. Recently, the transformer [16] has been applied to End-to-End speech recognition tasks [17], [18], [19], [20] and achieved promising results.

In this paper, we establish ASR systems for Lhasa dialect using the state-of-the-art the transformer-based End-to-End acoustic model. To use the low-resource data: We developed an effective model initialization method. Secondly, we discovered a set of highly compressed and reliable modeling units, which is first time used for Tibetan language speech recognition to our best knowledge. We also jointly training the transformer-based End-to-End acoustic model with two different acoustic unit sets. Finally, an error-correction dictionary is introduced to further improve the system performance.

The rest of this paper is organized as follows. The related works are overviewed in Section II. In Section III, the task data and the baseline systems of this paper are introduced. In Section IV, the proposed method for our task is explained and evaluated. This paper concludes in Section V.

II. Related Works

Several areas most related to our research are listed as follows.

A. Background knowledges of Lhasa Tibetan language

As we introduced in Section I. Tibetan language belongs to the Sino-Tibetan family. It has three dialects, including Lhasa Tibetan, Khams Tibetan, and Amdo Tibetan.

As shown in Figure 2, a typical Lhasa Tibetan character has a set of basic components: root-script (Root.), pre-script

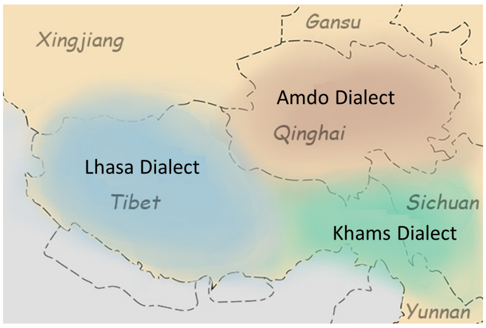
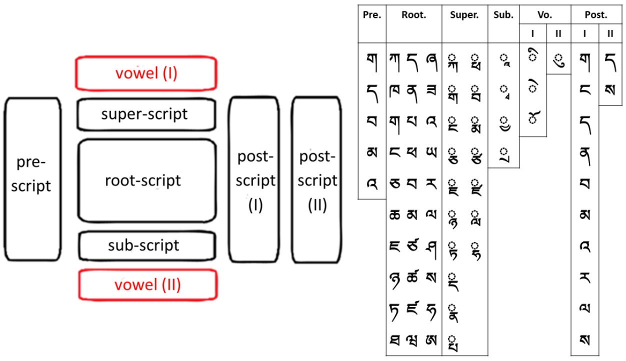


Fig. 1. Geological distribution of three major Tibetan dialects.

(Pre.), super-script (Super.), sub-script (Sub.), vowels (Vo.) and post-scripts (Post.) to express a wide range of grammatical categories and speech changes, e.g., number, tense and case, resulting in extremely large vocabularies.



Pre.	Root.	Super.	Sub.	Vo. I	Vo. II	Post. I	Post. II
ག	ཀ ད ཞ	◌ྐ ◌ྤ	◌ྱ	◌ི	◌ུ	ག	ད
ད	ཁ ན ཟ	◌ྒ ◌ྦ	◌ྲ	◌ེ		ང	ས
བ	ག པ འ	◌ྔ ◌ྨ	◌ླ	◌ོ		ད	
མ	ང ཕ ཡ	◌ྕ ◌ྩ	◌ྭ			ན	
འ	ཅ བ ར	◌ྗ ◌ྫ				བ	
	ཆ མ ལ	◌ྙ ◌ྷ				མ	
	ཇ ཙ ཤ	◌ྟ ◌ྟ				འ	
	ཉ ཚ ས	◌ྡ				ར	
	ཏ ཛ ཧ	◌ྣ				ལ	
	ཐ ཝ ཨ	◌ྤ				ས	

Fig. 2. The structure of a character in the Lhasa Tibetan writing systems.

As shown in Figure 2, the phone set can be defined differently by different combinations of these components, which will influence the speech recognition performance. And the initials come from the pronunciation of these components. For Lhasa dialect, the real initials are 28. Tibetan finals are determined by the possible combination of the character vowel and its post-scripts.

In our previous reseach [4], [3], we choose the initial/final based non-tonal phone set as acoustic modeling unit. The non-tonal phone set was built by referring to the previous phonological studies of Lhasa spoken language [21]. There are totally 29 initial consonants and 48 final units without considering the tones.

Since Lhasa Tibetan has no conclusive tonal pattern yet, a four-tone pattern is designed based on the four contour contrasts scheme [4]. The 48 non-tonal finals are extended to 192 tonal finals. The 29 initials are kept unchanged. Since the pitch-related features are still under-development, we only use the filter-bank feature in this paper.

B. State-of-the-art End-to-End ASR systems with transformer

Recently, the transformer-based model [16] has been successfully applied to various of ASR tasks [17], [20], [18] and showed promising results. The transformer-based model for ASR task maps an input speech feature sequenceto a sequence of intermediate representations in the encoder. Then, the decoder then generates an output sequence of symbols (phonemes, syllables, words, sub-words or words) given the intermediate representations. The biggest difference with those commonly used End-to-End models [9], [10] is the transformer-based acoustic model totally relies on no-recurrence components [16]: multi-head self-attention (MHA), positional-encoding (PE) and position-wise feed-forward networks (PFFN).

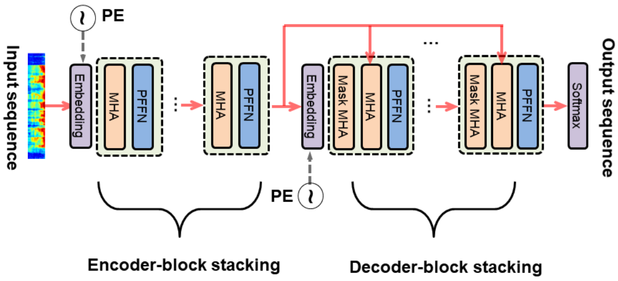


Fig. 3. The structure of transformer model.

As showin in Figure 3, for blocks in the encoders and decoders, they are defined as follows:

1. The encoder-block has MHA and PFFN layer consequently. Residual connections are used around each of the MHA and PFFN layers. Residual dropout [22] is introduced to each residual connection.

2. The decoder-block is similar to the encoder-block except inserting one MHA layer to perform attention over the output of the encoder-block stack.

3. PEs are added to the input at the bottoms of these encoder-block and decoder-block stacks, providing information about the relative or absolute position of the tokens in the sequence.

In this paper, we build End-to-End ASR systems for Lhasa dialect based on this transformer model.

III. Task Description and Baseline Systems

A. Data sets

Our speech corpus has 35.82 hours speech signal data which was collected from 23 Tibetan Lhasa native speakers, including 13 males and 10 females. All the speakers are college students whose mother tongue is Lhasa dialect. The speech signal is sampled at 16KHz 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 totally more than 38,700 sentences in the corpus.

TABLE I
SPEECH CORPUS OF LHASA DIALECT

Datasets	#Speakers	#Utterances	Hours
Training (**Lhasa-TRN**)	10M + 7F	36,090	31.9
Development (**Lhasa-DEV**)	3M + 3F	1,700	1.5
Testing (**Lhasa-TST**)	3M + 3F	2,664	2.4

B. Baseline system description

We train the baseline model using the 31.9 hours of training data (Lhasa-TRN). We first trained a GMM-HMM model using the MFCC feature with linear discriminant analysis (LDA), a maximum likelihood linear transform (MLLT) and feature space maximum likelihood linear regression (fMLLR)-based speaker adaptive training (SAT). We choose the initial/final based phone set (29 initals and 192 tonal finals) as acoustic modeling units following [4] and they clustered into 3320 tied-triphone states during training the GMM-HMM model.

Then, we train a DNN model with five hidden layers each comprising 2048 hidden nodes. The output layer had about 3320 nodes that corresponded to the tied-triphone states of the GMM-HMM model. We used 40-dim filter-bank features together with its 1st and 2nd order derivatives to train DNN model. All these features are mean and variance normalized (CMVN) per speaker. The filter-bank features of both the previous and subsequent five frames (11 frames of features in total) are added when inputting them into the DNNs. The DNN model is initialized using unsupervised pre-training and supervised fine-tuning using standard stochastic gradient descent (SGD) based on the cross-entropy loss criterion. The hyperparameters are adjusted based on the development set (Lhasa-DEV). All these were implemented using the Kaldi toolkit [23]. For testing, we decoded the sentences from test set (Lhasa-TST) with trigram character-based language model in the WFST decoding framework and evaluated our models using the character error rate (CER%). The ASR performance is 35.9% of CER%.

IV. THE LOW-RESOURCE LHASA DIALECT END-TO-END ASR SYSTEMS

A. Training baseline End-to-End ASR systems

We used the implementation of the transformer-based neural machine translation (NMT) [16] in tensor2tensor [1] for all our experiments. The training and testing settings are similar to [20] and they are listed in Table II.

TABLE II
MAJOR EXPERIMENTAL SETTINGS

Model structure	
Attention-heads	8
Hidden-units	512
Encoder-blocks	6
Decoder-blocks	6
Residual-drop	0.3
Attention-drop	0.0
Training	
Max-length	5000
Tokens/batch	10000
Epochs	30
Label-smooth	0.1
GPUs (K40m)	4
Warmup-steps	12000
Steps	300000
Optimizer	Adam
Testing	
Ave. chkpoints	last 20
Batch-size	100
Beam-size	13
Length-penalty	0.6
GPUs (K40m)	4

We used 120-dim filter-bank features (40-dim static $+\Delta$ $+\Delta\Delta$), which are mean and variance normalized per speaker, and 4 frames were spliced (3 left, 1 current and 0 right). Speed-perturbation [24] is not used to save training time. We trained the transformer-based acoustic models using the training set (Lhasa-TRN) of Lhasa dialect. We use 2072 characters as the basic acoustic units.

For testing, we decoded the sentences from test set (Lhasa-TST) without language model and evaluated our models using the character error rate (CER%).

However, the performance was rather poor (97.2% of CER% on Lhasa-TST) if the transformer-based acoustic model is trained with a random initialization from scratch. The reason for the poor performance could be the training data is too few but the parameters of the transformer-based acoustic model are relatively large (more than 200M) in this work. In next subsections, we introduce the proposed methods to effectively use the low-resource data.

B. Effective model initalization schemes

To compensate for the low-resourced training data, we proposed to use a well-trained transformer model to initialize our model. Its softmax layer is replaced by the language-specific softmax layer which is initialized randomly. All the existing languages in the world are so different in pronunciation, grammar, and syntax. We believe using the model well-trained from languages similar to Lhasa Tibetan dialect (e.g., Mandarin in Figure 4) can effectively initialize the model training process.

$$\begin{cases} Indo-European \mapsto \begin{cases} Germantic \mapsto English \\ Italic \mapsto Spanish, French, Italian \end{cases} \\ Sino-Tibetan \mapsto \textbf{Mandarin}, Myanmar, \textbf{Tibetan} \\ Altaic \mapsto Japanese, Korean \\ Austric \mapsto Vietnamese, Thai, Indonesian \end{cases}$$

Fig. 4. A brief summary of language family tree.

In this paper, we use a Mandarin transformer-based model (8 head-attention, 6 encoder-blocks and 6 decoder-blocks with

[1] https://github.com/tensorflow/tensor2tensor

512 nodes) trained from 178 hours of speech data selected from AIShell dataset [25] with the CER of 9.0%.

Through this initialization method, the transformer can converge very well. The CER% on Lhasa-TST reduced from 97.2% to 38.9%, which is very close to the ASR performance of DNN-HMM baseline (35.9%). Impressed by the effectiveness, we conduct this initialization method for all following experiments in the next subsection. However, when we use the same Mandarin transform-based model to initialize the Japanese model also using 31 hours data from CSJ corpus [26], there is only a little improvement from 55.1% CER% to 50.0%. It shows that selecting the language specified model for initialing training low-resource model should be worth investigation.

We also observed that we can't initialize the target system (filter-bank) using a transformer-based model trained from a different feature (filter-bank with pitch feature [27]). The CER% on Lhasa-TST had a sharp increase to 59.4%. It means that this method is feature dependent.

C. The highly compressed sub-character units

A Tibetan character is further segmented to a sequence of sub-character tokens as shown in Figure 5. The vertically stacking components (super-script, root-script, sub-script and vowels) in a character are seperated and regarded as individual units. The boundary mark between two successive characters is also regarded as an individual unit. We name this sub-character unit set "basic-57." We got confirmations from the linguists that the original characters can be easily recovered as long as the boundary marks exist.

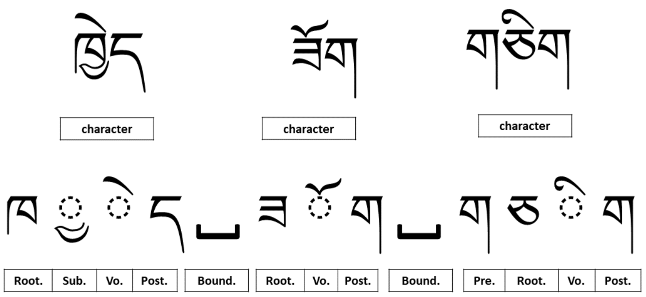


Fig. 5. Segmenting characters to sub-character units.

Several modeling units were compared on Lhasa dialect ASR tasks, including phones, characters and the sub-characters. As alternatives to the basic-57, the word-piece-model (WPM)[28] is also used to segment the characters with predefined unit number (100, 300, 500, 700, and 900). We used the sentence-piece toolkit [2] as the sub-character segmenter. We compare the ASR performances of following transformer systems modeled with different acoustic units with the DNN-HMM system as shown in Table III.

TABLE III
ASR PERFORMANCE (CER%) OF TRANSFORMER-BASED MODELS TRAINED WITH DIFFERENT UNITS COMPARED WITH THE PHONE-BASED DNN-HMM BASELINE SYSTEM

Network	#unit	CER% on Lhasa-TST
Phone DNN-HMM	Senone 3320	35.9%
Char. transformer	Char. 2072	38.9%
Sub-char. transformer	Basic 57	37.3%
	WPM 100	40.5%
	WPM 300	39.4%
	WPM 500	39.8%
	WPM 700	40.7%
	WPM 900	39.4%
Multi-unit transformer	Basic 57+Char. 2072	36.3%
+ error-correction dictionary		35.3%

The model trained with the basic-57 unit (**Sub-char. transformer**) achieved the closest performance with the baseline (**DNN-HMM**) model, and significantly (two-tailed t-test at p-value $<$ 0.05) outperformed the character-based model (**Char. transformer**). The transformer models trained with other sub-character unit sets generated by WPM model can't outperform the basic-57 based transformer. The small performance gap between the basic-57 based transformer and baseline DNN-HMM system can be compensated with the acceleration on the training and decoding speed. The acceleration comes from two parts: the first part is the simplified training and decoding schemes by using End-to-End training. The other part comes from the highly compressed sub-character-based acoustic units, which shows its reliability and advantages compared with previous units.

D. Multi-unit Training and Error-correction Dictionary

In our experiment, we also find that joint training the transformer model with two different units (Basic 57 and Char. 2072) together using the multilingual training method described in [18]. This model (**Multi-unit transformer**) can significant improve the system performance of Sub-char. transformer. Further improvement can be achieved by introducing an error-correction dictionary, which is statistically generated by comparing the recognition result and oracle on training data using SCTK. Based on these two techniques, our proposed Sub-char. units can make the transformer-based End-to-End ASR system outperform the **DNN-HMM** baseline model.

V. Conclusion

In this paper, we focus on training transformer-based End-to-End ASR systems for Lhasa dialect. To effectively making use of the low-resource data, we investigate effective initalization strategies, a compressed acoustic modeling unit set, multi-unit training and error-correction dictionary. Experiments show our proposed method can effectively modeling low-resource Lhasa dialect and outperforms conventional DNN-HMM baselines. We believe that our work will promote the existing speech recognition research on Tibetan language.

[2] https://github.com/google/sentencepiece

Acknowledgment

The authors thank Dr. Hongcui Wang for building concrete baseline systems and Mr. Kuntharrgyal Khyuru for discussions about Tibetan language and text processing.

References

[1] L. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1988.

[2] G. Dahl, D. Yu, L. Deng, and A. Acero, "Context dependent pre-trained deep neural networks for large vocabulary speech recognition," *IEEE Trans. ASLP*, vol. 20, no. 1, pp. 30–42, 2012.

[3] H. Wang, K. Khyuru, J. Li, G. Li, J. Dang, and L. Huang, "Investigation on acoustic modeling with different phoneme set for continuous Lhasa Tibetan recognition based on DNN method," in *Proc. APSIPA ASC*, 2016.

[4] J. Li, H. Wang, L. Wang, J. Dang, K. Khyuru, and G. Lobsang, "Exploring tonal information for lhasa dialect acoustic modeling," in *Proc. ISCSLP*, 2016.

[5] D. Povey, V. Peddinti, D. Galvez, P. Ghahrmani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for asr based on lattice-free MMI," in *Proc. INTERSPEECH*, 2016.

[6] J. Yan, Z. Lv, S. Huang, and H. Yu, "Low-resource Tibetan dialect acoustic modeling based on transfer learning," in *Proc. The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages*, 2018.

[7] A. Graves and N. Jaitly, "Towards End-to-End speech recognition with recurrent neural networks," in *Proc. ICML*, 2014.

[8] Y. Miao, M. Gowayyed, and F. Metze, "EESEN: End-to-End speech recognition using deep RNN models and WFST-based decoding," in *Proc. IEEE-ASRU*, 2015, pp. 167–174.

[9] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Proc. NIPS*, 2015.

[10] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proc. IEEE-ICASSP*, 2016.

[11] H. Hadian, H. Sameti, D. Povey, and S. Khudanpur, "End-to-end speech recognition using lattice-free mmi," in *Proc. INTERSPEECH*, 2018.

[12] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid CTC/attention architecture for end-to-end speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.

[13] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "Espnet: End-to-end speech processing toolkit," in *Proc. INTERSPEECH*, 2018.

[14] S.Ueno, H.Inaguma, M.Mimura, and T.Kawahara, "Acoustic-to-word attention-based model complemented with character-level ctc-based model," in *Proc. IEEE-ICASSP*, 2018.

[15] T. Hori, S. Watanabe, Y. Zhang, and W. Chan, "Advances in joint CTC-Attention based End-to-End speech recognition with a deep CNN Encoder and RNN-LM," in *Proc. INTERSPEECH*, 2017.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *arXiv preprint arxiv:1706.03762*, 2017.

[17] L. Dong, S. Xu, and B. Xu, "Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition," in *Proc. IEEE-ICASSP*, 2018.

[18] S. Zhou, S. Xu, and B. Xu, "Multilingual end-to-end speech recognition with a single transformer on low-resource languages," in *arXiv preprint arxiv:1806.05059*, 2018.

[19] S. Zhou, L. Dong, S. Xu, and B. Xu, "A comparison of modeling units in sequence-to-sequence speech recognition with the transformer on mandarin chinese," in *arXiv preprint arxiv:1805.06239*, 2018.

[20] S. Zhou, L. Dong, S. Xu, and B. Xu, "Syllable-based sequence-to-sequence speech recognition with the transformer in mandarin chinese," in *Proc. INTERSPEECH*, 2018.

[21] *Tibeto-Chinese Lhasa Vernacular Dictionary (Tibetan)*, The Ethnic Publishing House, 1983.

[22] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Weinberger, "Deep networks with stochastic depth," in *Proc. ECCV*, 2016.

[23] D. Povey and et al., "The Kaldi speech recognition toolkit," in *Proc. IEEE-ASRU*, 2011.

[24] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Proc. INTERSPEECH*, 2015.

[25] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "AIShell-1: An open-source Mandarin speech corpus and a speech recognition baseline," in *Proc. Oriental COCOSDA*, 2017.

[26] K. Maekawa, "Corpus of spontaneous japanese: Its design and evaluation," in *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003.

[27] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal, and S. Khudanpur, "A pitch extraction algorithm tuned for automatic speech recognition," in *Proc. IEEE-ICASSP*, 2014.

[28] T. Kudo, "Subword regularization: Improving neural network translation models with multiple subword candidates," in *arXiv preprint arxiv:1804.10959*, 2018.