Enriching Under-Represented Named Entities for Improved Speech Recognition

Tingzhi Mao* Yerbolat Khassanov†§, Van Tung Pham†, Haihua Xu†,
Hao Huang*, Aishan Wumaier* and Eng Siong Chng†
* School of Information Science and Engineering, Xinjiang University, Urumqi, China
† School of Computer Science and Engineering, Nanyang Technological University, Singapore
§ ISSAI, Nazarbayev University, Kazakhstan

***Abstract*—Automatic speech recognition (ASR) for under-represented named-entity (UR-NE) is challenging due to such named-entities (NE) have insufficient instances and poor contextual coverage in the training data to learn reliable estimates and representations[1]. In this paper, we propose approaches to enriching UR-NEs to improve speech recognition performance. Specifically, our first priority is to ensure those UR-NEs to appear in the word lattice if there is any. To this end, we employ class-based language model (LM) philosophy, making exemplar utterances for those UR-NEs according to their classes (e.g. location, person, organization, etc.), ending up with an improved LM that boosts the UR-NE occurrence in the word lattice. Then we boost the recognition performance through lattice rescoring methods. We first enrich the representations of UR-NEs in a pretrained recurrent neural network LM (RNNLM) by borrowing the embedding representations of the rich-represented NEs (RR-NEs), yielding the lattices that statistically favor the UR-NEs. Finally, we directly boost the likelihood scores of the utterances containing UR-NEs and gain further performance improvement.**

I. Introduction

With the surge of voice-enabled applications in smart devices, the correct recognition of under-represented named-entity (UR-NE) became vital, especially for the downstream applications that employ ASR outputs [1], [2], [3], [4]. Since the UR-NEs might constitute the essential details of an utterance such as person, location, and organization names.

However, speech recognition for UR-NE words is challenging. This is because those UR-NEs rarely occur in the training corpus, and they also lack of contextual information, resulting in weaker acoustic and language models on such NEs.

To boost the recognition performance of those UR-NEs, our first priority is to ensure that they appear in a decoded lattice. Then we propose a series of approaches to extract the final one-best hypothesis that contains UR-NEs if there is any.

Specifically, we propose a simplified class-based language modeling framework that is aimed to boost the n-gram count for those UR-NEs under word-based n-gram langauge model (LM) context. In practice, we first define NE class, but we actually avoid building NE-class-based n-gram LM, since it is not the NE-class-based n-gram LM, but the word-based n-gram LM instead, that is employable to generate word lattice. Besides, class-based to word-based n-gram LM expansion is not a trivial work. It entails big memory and post-pruning on the generated word-based n-gram LM. As a result, we propose to generate exemplar utterances for the UR-NEs to boost word-based n-gram LM directly according to the defined NE classes.

We found that the UR-NE occurrence in the decoded lattice significantly improved with the help of UR-NE enriched n-gram LM. After that, we first rescore lattice by a improved UR-NE-enriched recurrent neural language model (RNNLM) [5] that has proved its effectiveness for UR-NEs recognition. lastly, we directly boost the utterances that contain UR-NEs in lattices. The combination of these approaches significantly improves the recognition performance of UR-NEs

The paper is organized as follows. Section II briefs the prior related work. In Section III, we are briefing data specification for the experiment. We then detail the proposed approaches to enriching representations for UR-NEs in Section IV. Section V is for our experimental setup and results. We draw conclusions in Section VI.

II. Related work

In ASR community, a common practice is " the more data, the better". However, for those UR-NEs, even huge data cannot guarantee they are covered or sufficiently covered that leads to insufficient contextual information. Prior work is mostly focused on out-of-vocabulary (OOV) words [6], [7], [8], [9]. However, performance would be sub-optimal by simply adding those OOV words to the ASR dictionary. This is because those OOV words have no context information, which is essential to the decode utterances containing OOV words. some people propose class-based LM [10], [11], [12], [13], [14], [15], among them [10], [11], [12], [13] are class-based ngram LM that is troublesome, and [14], [15] these can reduce the cost but they are weaker performance than word-based RNNLM. Another effort direction on rare word recognition is to use diversified LMs [5], [16], [17], [18], [19], to rescore lattice/N-best hypotheses, hopefully to achieve improved results. For example, [17], [18], [19] proposed to employ hybrid word/subword tokens as input and output units of neural LMs. On the other hand, [5], [16] proposed to augment the representations of rare words in embedding

* Tingzhi Mao is an intern at MICL Lab, NTU, Singapore. Hao Huang is the corresponding authors

[1]In this paper, UR-NE refers to the named-entity (NE) words that have low-frequency count, say, the count is in [1, 9] in this work, or do not appear in the training data at all, i.e. the count is 0.

matrices of pretrained word-level neural LM. The success of the above-mentioned methods relies on a presumption that those rare words do appear in the lattice which are often generated with the help of a cheaper N-gram LM. However, such assumption is not always guaranteed.

In this work, to conduct Singapore street NE recognition, we do not rely on much extra text or acoustic data. Instead, we fully exploit our training data to discover the pattern of those utterances containing NEs that are rich-represented (RR). We employ those RR-NE utterances as pool to generate exemplar utterances for those UR-NEs according to the NE category correspondence. We demonstrate such an N-gram LM boosted with the exemplar utterances can be significantly helpful on the UR-NE occurrence in the resulting lattice. We further apply rescoring methods on the boosted lattice to achieve improved results.

III. Data Specification

To train the ASR, we utilize National Speech Corpus (NSC) [20] developed to advance the english speech related technologies in Singapore. The NSC consists of three main parts: 1) read speech using phonetically balanced scripts, 2) read speech featuring words pertinent to the Singapore context, containing a lot of NEs, and 3) conversational speech. To evaluate the proposed approaches, we use the subset of the third part (NSC-part3) as training data and a small portion of the second part (NSC-part2) as an evaluations set. In addition, we also use SG-streets[2] data set as an additional evaluation set. The SG-streets data set consists of six recordings where Singaporean students read English passages containing Singapore street NEs. The detailed data description is shown in Table I. From Table I, we can see the two test sets contain a lot of NEs. This is particularly true for the NSC-part2 test set. It contains a lot of sentences with NEs, for instance, utterances like "`please look for Makaila when you reach Kallang wave mall`". Besides, the OOV rate related to the training data is also quite high for both test sets. This is because the training data is conversational data, and the vocabulary is usually small (∼20k in our case).

TABLE I: The overall data set specification

Category	Train	Test	
		NSC-part2	SG-streets
Speakers	482	76	6
Duration (hrs)	100	1.6	1.0
Utterances	137,058	1,176	517
NE rate (%)	0.62	19.33	7.82
OOV rate (%)	-	12.93	8.25

IV. Approaches to Enriching UR-NE

A. Enrich UR-NEs with exemplar utterance for Language Modeling

As mentioned, UR-NEs refer to the NEs whose count falls in, say, [0, 9], in the training data. Here, zero count means out-of-vocabulary words. As indicated in Table I, the two test sets not only contain high NE rate but also contain high OOV rate; and most of the OOV words belong to NEs. Specifically, the NEs mainly falls in 6 class, that is, location, person, country, company, organization, and city. Linguistically, the NEs are mostly from Mandarin, Cantonese, Hokkien, and Malay languages.

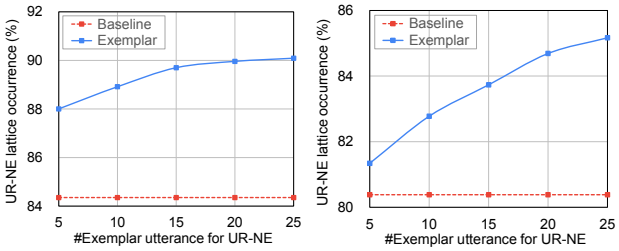
Simply adding all those UR-NEs into the ASR lexicon is a natural choice, however, recognition improvement won't be fully unleashed. This is because the UR-NEs still lack of linguistic context that is important for LM. To address this problem, we propose a simplified class-based LM that is to generate exemplar utterances for them and merge such exemplar utterances to boost the language model.

In practice, we manually classify the NEs in our training data into two parts according to their count. We name those NEs with higher count (say, count in $[10, +\infty)$) as RR-NE. Simultaneously, we label the NEs with 6 class respectively. We then build `hash tables` for the NEs and corresponding exemplar utterances that are randomly selected from the training transcript.

Given those UR-NEs, we label them with one of the 6 classes as mentioned above, we then search the class by looking-up the `hash tables` we built, yielding exemplar utterances for those RR-NEs. By simply substitute RR-NEs with the UR-NEs, we obtain corresponding exemplar utterances for each of UR-NEs. Finally, we use this class-based N-gram LM to decode.

Figure 1 illustrates the effectiveness of using exemplar utterance on UR-NE occurrence in the lattices for the 2 test sets in Table I. As is clearly shown in Figure 1, even with 5 exemplar utterances for UR-NE can lead to significant UR-NE occurrence in the lattices.

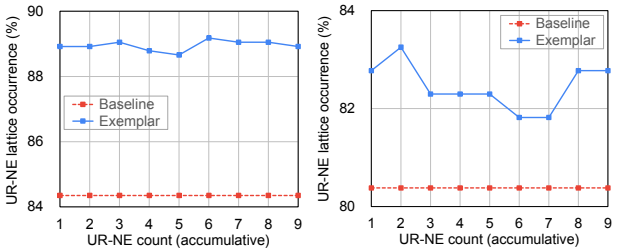
Meanwhile, we are also curious about on which part of UR-NE the exemplar method has most impact. To do this, we divide the UR-NEs into several groups according to their counts in training data. Specifically, the count of the overall NEs we are considering lies in [0, 9], we divide them into 9 groups, we then analyze the NE lattice occurrence for each sub-group accumulatively, that is, each sub-group [0,b] means



(a) NSC-part2 (b) SG-streets

Fig. 1: Effectiveness of using exemplar utterances on the UR-NE occurrence in lattices. Here, UR-NE count falls in [0, 10) in training data, and all test NEs are included in ASR lexicon for the baseline which corresponds to 0 exemplar utterances.

[2]https://github.com/khassanoff/SG_streets



(a) NSC-part2 (b) SG-streets

Fig. 2: NE occurrence in lattices versus the upperbound count of UR-NE. Here the denominator is the number of NEs whose frequency counts are in [0, 9].

we consider the NEs with count in the corresponding level, and $b \in [0, 9]$. Figure 2 plots the NE lattice occurrence versus NE counts in training data. As is observed from Figure 2, the exemplar method only has significant impact on those extremely under-represented NEs in training data, specifically, whose count lies in [0,1], namely, out-of-vocabulary and single count NEs. Thirdly, We also analyzed the number of RR-NEs and the number of utterances per RR-NEs and find that this is not a very important parameter to care.

Based on the above analysis on the proposed exemplar utterances method, we use following settings for the experiments in Section V. We randomly select 20 RR-NEs from the training data, After that, we only boost those UR-NEs whose counts are in [0, 1] in training data, and for each UR-NE, we randomly generate 10 exemplar utterances from the utterance pool.

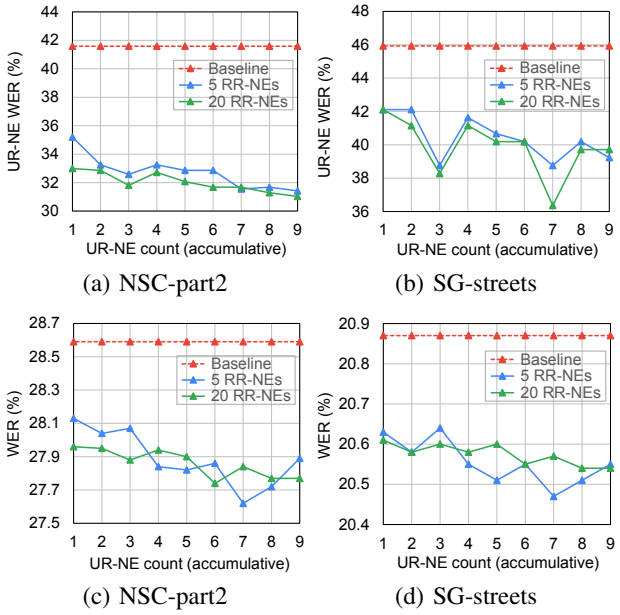
B. Enrich UR-NE for lattice rescoring

1) UR-NE-enriched RNNLM lattice rescoring: RNNLM lattice rescoring is an effective approach to boosting ASR performance as a post-processing strategy [19], [21]. To extract better results from lattice in Section IV-A, we propose an improved RNNLM that enrich the representations for those UR-NEs following the prior work [5]. Specifically, we enrich the embedding vectors of UR-NEs in space using embedding vectors of the RR-NEs that are corresponding set of similar class words. while we keep the parameters of pre-trained RNNLM unchanged. The formula is as follows:

$$\hat{e}_u = \frac{e_u + \sum_{e_c \in \mathcal{C}_r} m_c e_c}{|\mathcal{C}_r| + 1} \quad (1)$$

where C_r is the overall embedding vector of RR-NEs set, e_u is UR-NE embedding and $\hat{e}_u$ is the enriched representation of e_u. m_c is a metric to weigh the relevance of RR-NEs to the corresponding UR-NE. In this paper, m_c is 0.7 if two NEs are in the same class, and 0.3 for the other case.

For the proposed method (for simplicity from now on, we also name it as `RNNLM-enriched` method), the success of Equation 1 is dependent on 2 factors, that is, how many RR-NEs and UR-NEs are involved. Besides, we also prefer the overall WER and NE-WER improvement is positively correlated. Figure 3 shows both NE-WER and WER curves



(a) NSC-part2 (b) SG-streets

(c) NSC-part2 (d) SG-streets

Fig. 3: UR-NE WER (%) and overall WER (%) versus different UR-NEs that are enriched with Equation 1.

versus UR-NEs that are defined by their count in training data, with different RR-NE embeddings that are employed in Equation 1. From Figure 3, we see that by using very small RR-NE embeddings (5 RR-NEs) to enrich those UR-NEs whose count is ∼7, both NE-WER and WER are consistently improved. We note that the baselines here are the corresponding normal RNNLM resoring results. More importantly, the definitions of UR-NEs and RR-NEs here are separated with corresponding definitions in Section IV-A, that is, both methods can choose completely different UR-NEs and RR-NEs to enrich. In Section V, we always choose 5 RR-NE embeddings to enrich those UR-NEs whose count in [0,7].

2) UR-NE-biased lattice rescoring: Even with the rescored lattice by the RNNLM-enriched method, it cannot be guaranteed the UR-NEs to appear in the 1-best hypothesis of the utterances in the end. This is because the hypothesis containing UR-NEs might have too low likelihood scores. Therefore, we propose a UR-NE-biased lattice rescoring method (For simplicity, we also call it as `Lattice boosting` method) to boost the hypothesis containing UR-NEs.

Practically, we are doing what follows in order. We treat all UR-NEs as keywords. We build a trivial finite-state-transducer (FST) [22], [23] for the entire UR-NE set. For a given word lattice, we first perform keyword search [24] by composing the UR-NE FST with the transformed lattice. This determines if any UR-NE exists and corresponding location in the lattice. If there is UR-NE, we extract the best path/hypothesis containing the UR-NE. It is simply a forward-backward path search operation on the lattice.

TABLE II: The overall WERs (%) and NE-WERs (%) with the proposed methods. The NE-WER is computed only for the NEs whose count is in [0, 9]. Here, zero mean the NEs are absent from the training data, while [1,9] refers to the rare case. "Lattice boosting" means we perform UR-NE-biased lattice rescoring in Section IV-B2.

ID	System	NSC-part2				SG-streets			
		WER (%)	NE-WER (%)			WER (%)	NE-WER (%)		
			Rare	Absent	ALL		Rare	Absent	ALL
S1	Baseline	30.14	30.81	46.05	42.37	22.12	18.57	58.99	45.45
S2	S1 + Exemplar utterance	27.35	28.11	36.43	34.42	19.85	15.71	56.83	43.06
S3	S2 + RNNLM	26.80	28.11	35.22	33.51	19.41	18.57	55.40	43.06
S4	S2 + RNNLM-enriched	**26.67**	26.49	32.65	31.16	**19.35**	17.14	50.36	39.23
S5	S4 + Lattice boosting	29.88	**11.89**	**20.79**	**18.64**	23.04	7.14	**30.94**	**22.97**
S6	S1 + RNNLM	28.59	30.27	45.19	41.59	20.87	21.43	58.27	45.93
S7	S1 + RNNLM-enriched	27.62	24.86	33.68	31.55	20.47	14.29	51.08	38.76
S8	S7 + Lattice boosting	29.69	12.97	23.54	20.99	22.25	**4.29**	33.81	23.92

V. Experimental setup and results

Experiments are conducted with `Kaldi` toolkit [3]. The acoustic models are the factorized time-delay neural network (TDNN-F) [25], trained with lattice-free maximum mutual information (LF-MMI) [26] criterion. All are a 6-layer convolutional neural network topped with 11-layer TDNN-F network with each layer having 1536 input neurons and 256 bottleneck output neurons respectively. The ASR lexicon is position-dependent grapheme lexicon [27], and the vocabulary is 21.7k. Our grapheme lexicon can make comparable results with the conventional phonetic lexicon on those in-vocabulary words, while it yields better results on the NE recognition. To realize OOV-free on either test set, we collect $\sim$3000 NEs from website[4], which covers the NE-related OOV words on either test set. We use 4-gram LM trained with training transcript to perform first-pass decoding to generate lattice. After that, we employ RNNLM [19] to conduct lattice rescoring. To achieve a desired performance, we also employ both speed perturbation [28], [29], as well as SpecAugment [30] simultaneously.

Table II presents the overall results with (S2-S5) or without (S6-S8) exemplar method in Section IV-A employed. From Table II, exemplar method achieves better results on the final UR-NE recognition, comparing S5 with S8. On NSC-part2, the overall NE-WERs are 18.64% versus 20.99%; the NE-WERs are 22.97% versus 23.92% on SG-streets. It is particularly effective on the NEs that are absent in the training data, yielding consistent improved results on either test set. One point worth a notice is that it seems the exemplar method is not perfectly coordinated with the RNNLM-enriched method. Comparing S3 with S6, exemplar method yields obvious better results, however, after the RNNLM-enriched method is applied, the benefit is significantly reduced. As is seen from S4 versus S7, exemplar method even yields worse results (17.14% versus 14.29%) on SG-streets "Rare" case. We conjecture it diverts the embedding e_u in Equation 1. Besides, taking a closer look into the data, we found the sentences of the SG-streets have few repetitive patterns, and they are much longer. For the NSC-part2, the patterns of the utterances containing NEs are rather restricted and repetitive, which favors for the exemplar method, since it is much easier to capture the limited context.

Table II also shows the effectiveness of the RNNLM-enriched method on the overall WER and NE-WER improvement, with or without exemplar utterance method, as can be clearly observed in S4 and S7. Additionally, by lattice boosting method, we can obtain remarkable NE-WER reduction on both data sets. For instance, the NE-WER is down to 22.97% from 39.23% in exemplar case, and 23.92% from 38.76% without exemplar method on SG-streets test set. However, the overall WERs are slightly worsen, as is seen when S4 versus S5, and S7 versus S8 are respectively compared.

Finally in Table II as mentioned, we observe RNNLM-enriched method is not well coordinated with the exemplar method. We guess this is because the exemplar utterances might divert the embedding estimate for the NR-NEs in Equation 1. To verify our conjecture, we decouple the two methods, that is, we use exemplar method to generate lattice, but stick with using original training transcript to train RNNLM and let the afterwards RNNLM-enriched method not be affected by the exemplar method. Table III reports the NE-WER results. Compared with results in Table II, we notice that we make the best NE-WER results on either test sets.

TABLE III: NE-WER results of Decoupled exemplar and RNNLM-enriched methods. S9 stands for lattice rescoring that corresponds to S3 and S6 in Table II, S10 refers to RNNLM-enriched method, corresponding to S4 and S7 in Table II, S11 refers to lattice-boosting method, whose counterpart is S5 and S8 in Table II.

ID	NE-WER (%)					
	NSC-part2			SG-streets		
	Rare	Absent	ALL	Rare	Absent	ALL
S9	27.03	36.60	34.29	15.71	52.52	40.19
S10	24.32	32.30	30.38	15.71	47.48	36.84
S11	12.43	20.62	18.64	4.29	30.94	22.01

[3]https://github.com/kaldi-asr/kaldi

[4]https://geographic.org/streetview/singapore

VI. Conclusions

In this paper, we proposed a bunch of approaches to enrich under-represented name-entities, yielding better name-entity recognition performance. To realize this, our first objective is to guarantee the occurrence of the under-represented name-entity is improved in decoded lattice. Consequently we introduce an exemplar utterance generation method, yielding an improved n-gram LM that favors for the under-represented name-entities. Though the exemplar method is rather heuristic, we demonstrated its effectiveness. To achieve better results on the improved lattice, we then employed two lattice rescoring methods. One is the RNNLM-enriched lattice rescoring method, by enriching embedding of the under-represented name-entity with corresponding embeddings of rich-represented name-entities. We found it is very effective to boost the overall ASR performance, that is, with or without name-entity recognition considered. Another method is we directly favor the utterance that contains under-represented name-entities from lattice. With such a method, we obtain slightly degraded ASR results, but significantly better results on the under-represented name-entity recognition.

Acknowledgment

This research is supported by the National Key R&D Program of China (2020AAA0107902); Opening Project of Key Laboratory of Xinjiang Uyghur Autonomous Region, China (2020D04047); Natural Science Foundation of China (62137002, 61663044, 61761041).

References

[1] M. Siu, T. Vessenes, I. Bulyko, and O. Kimball, "Improved named entity extraction from conversational speech with language model adaptation," in *2010 IEEE Spoken Language Technology Workshop*, 2010, pp. 418–423.

[2] C. Peyser, S. Mavandadi, T. Sainath, J. Apfel, R. Pang, and S. Kumar, "Improving tail performance of a deliberation e2e asr model using a large text corpus," in *INTERSPEECH*, 2020.

[3] C. Peyser, T. N. Sainath, and G. Pundak, "Improving proper noun recognition in end-to-end asr by customization of the mwer loss criterion," in *ICASSP*, 2020, pp. 7789–7793.

[4] F. Lux and N. T. Vu, "Meta-learning for improving rare word recognition in end-to-end asr," *ArXiv*, vol. abs/2102.12624, 2021.

[5] Y. Khassanov, Z. Zeng, V. T. Pham, H. Xu, and E. S. Chng, "Enriching Rare Word Representations in Neural Language Models by Embedding Matrix Augmentation," in *Interspeech*, 2019, pp. 3505–3509.

[6] Y. He, B. Hutchinson, P. Baumann, M. Ostendorf, E. Fosler-Lussier, and J. Pierrehumbert, "Subword-based modeling for handling oov words in keyword spotting," in *ICASSP*, 2014, pp. 7864–7868.

[7] E. Egorova and L. Burget, "Out-of-vocabulary word recovery using fst-based subword unit clustering in a hybrid asr system," in *ICASSP*, 2018, pp. 5919–5923.

[8] S. Thomas, K. Audhkhasi, Z. Tüske, Y. Huang, and M. Picheny, "Detection and recovery of OOVs for improved english broadcast news captioning," in *Interspeech*, 2019, pp. 2973–2977.

[9] X. Zhang, D. Povey, and S. Khudanpur, "Oov recovery with efficient 2nd pass decoding and open-vocabulary word-level rnnlm rescoring for hybrid asr," in *ICASSP*, 2020, pp. 6334–6338.

[10] W. Ward and S. Issar, "A class based language model for speech recognition," in *ICASSP 1996*, vol. 1, 1996, pp. 416–418 vol. 1.

[11] C. Samuelsson and W. Reichl, "A class-based language model for large-vocabulary speech recognition extracted from part-of-speech statistics," in *ICASSP 1999 (Cat. No.99CH36258)*, vol. 1, 1999, pp. 537–540 vol.1.

[12] J. Hoidekr, J. Psutka, and A. Praák, "Benefit of a class-based language model for real-time closed-captioning of tv ice-hockey commentaries," in *LREC*, 2006.

[13] W. Naptali, M. Tsuchiya, and S. Nakagawa, "Class-based n-gram language model for new words using out-of-vocabulary to in-vocabulary similarity," *IEICE Trans. Inf. Syst.*, vol. 95-D, pp. 2308–2317, 2012.

[14] Y. Shi, W. Zhang, J. Liu, and M. T. Johnson, "Rnn language model with word clustering and class-based output layer," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2013, pp. 1–7, 2013.

[15] I. S. Kipyatkova and A. Karpov, "Class-based lstm russian language model with linguistic information," in *LREC*, 2020.

[16] Y. Khassanov and E. S. Chng, "Unsupervised and efficient vocabulary expansion for recurrent neural network language models in asr," *Interspeech*, pp. 3343–3347, 2018.

[17] T. Mikolov, I. Sutskever, A. Deoras, H.-S. Le, S. Kombrink, and J. Cernocky, "Subword language modeling with neural networks," *preprint (http://www. fit. vutbr. cz/imikolov/rnnlm/char. pdf)*, vol. 8, 2012.

[18] Y. Kim, Y. Jernite, D. A. Sontag, and A. M. Rush, "Character-aware neural language models," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. AAAI Press, 2016, pp. 2741–2749.

[19] H. Xu, K. Li, Y. Wang, J. Wang, S. Kang, X. Chen, D. Povey, and S. Khudanpur, "Neural network language modeling with letter-based features and importance sampling," in *ICASSP*, 2018, pp. 6109–6113.

[20] J. X. Koh, A. Mislan, K. Khoo, B. U. Ang, W. Ang, C. Ng, and Y.-Y. Tan, "Building the singapore english national speech corpus," in *INTERSPEECH*, 2019.

[21] H. Xu, T. Chen, D. Gao, Y. Wang, K. Li, N. Goel, Y. Carmiel, D. Povey, and S. Khudanpur, "A pruned rnnlm lattice-rescoring algorithm for automatic speech recognition," *ICASS)*, pp. 5929–5933, 2018.

[22] C. Allauzen, M. Riley, J. Schalkwyk, W. Skut, and M. Mohri, "Openfst: A general and efficient weighted finite-state transducer library," in *CIAA*, 2007.

[23] M. Riley, C. Allauzen, and M. Jansche, "Openfst: An open-source, weighted finite-state transducer library and its applications to speech and language," in *HLT-NAACL*, 2009.

[24] D. Can and M. Saraclar, "Lattice indexing for spoken term detection," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 8, pp. 2338–2347, 2011.

[25] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, and S. Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks," in *INTERSPEECH*, 2018.

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

[27] D. Le, X. Zhang, W. Zheng, C. Fügen, G. Zweig, and M. L. Seltzer, "From senones to chenones: Tied context-dependent graphemes for hybrid speech recognition," in *ASRU*, 2019, pp. 457–464.

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

[29] V. Peddinti, G. Chen, V. Manohar, T. Ko, D. Povey, and S. Khudanpur, "Jhu aspire system: Robust lvcsr with tdnns, ivector adaptation and rnn-lms," in *ASRU*, 2015, pp. 539–546.

[30] D. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," in *INTERSPEECH*, 2019.