INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Learning Neural Network Representations using Cross-lingual Bottleneck Features with Word-pair Information

Yougen Yuan[1,2], *Cheung-Chi Leung*[2], *Lei Xie*[1], *Bin Ma*[2], *Haizhou Li*[2]

[1]Shaanxi Provincial Key Laboratory of Speech and Image Information Processing, School of Computer Science, Northwestern Polytechnical University, Xi'an, China
[2]Institute for Infocomm Research, A*STAR, Singapore
{ygyuan,lxie}@nwpu-aslp.org, {ccleung,mabin,hli}@i2r.a-star.edu.sg

Abstract

We assume that only word pairs identified by human are available in a low-resource target language. The word pairs are parameterized by a bottleneck feature (BNF) extractor that is trained using transcribed data in a high-resource language. The cross-lingual BNFs of the word pairs are used for training another neural network to generate a new feature representation in the target language. Pairwise learning of frame-level and word-level feature representations are investigated. Our proposed feature representations were evaluated in a word discrimination task on the Switchboard telephone speech corpus. Our learned features could bring 27.5% relative improvement over the previously best reported result on the task.

Index Terms: feature representations, pairwise learning, low-resource speech processing, bottleneck features (BNFs), Siamese network

1. Introduction

Deep learning has shown success and is widely used in acoustic modeling [1, 2, 3, 4]. Training a deep neural network usually requires a large amount of data together with their frame-level labels derived from word-level transcription and a pronunciation dictionary. This method cannot be applied to many languages in the world, especially low-resource languages and dialects.

Training neural networks (NNs) with paired examples has been proposed for various tasks [5, 6, 7]. For languages without any prior linguistic knowledge, it is difficult to give utterances with appropriate labels. However, it is easy to specify whether the words spoken in two utterances are the same. In [8, 9, 10], they use unsupervised term discovery [11] or transcriptions to find the same unknown type of word pairs as weak supervision, and train a deep architecture with this weak supervision to obtain better feature representations for a word or triphone discrimination task.

Recently bottleneck-type NNs trained using high-resource languages have been commonly used as a feature extractor for different tasks in low-resource languages. The corresponding bottleneck features (BNFs) form a compact (low-dimensional) representation capturing information for phone classification. Experiments in automatic speech recognition (ASR) have showed that the extractor of BNFs [12, 13] (and related MLP posteriors [14, 15] trained on large amounts of data could help to improve the recognition of a new target language. Moreover, cross-lingual BNFs have been widely used in language-independent query-by-example spoken term detection [16, 17, 18, 19, 20]. These works imply that using cross-lingual resource is a potential way to learn NN representations for deep architecture with limited resource in target language, and cross-lingual BNFs might be good feature representations for other languages.

In this paper, the cross-lingual portability of BNFs motivated us to perform pairwise supervision of NNs using cross-lingual BNFs of word pairs, and the new feature representations for the target language were generated from the pairwise supervised NNs. The learned NN representations were evaluated in a word discrimination task on the Switchboard telephone speech corpus. To our best knowledge, this study is the first attempt to use cross-lingual BNFs in pairwise supervision of NNs for this task. The NN feature representations learned in previous studies can be classified into frame-level and word-level feature representations, and our proposed method is feasible for learning these two types of feature representations. Our ultimate goal is to use the learned feature representations for downstream ASR and search tasks. Our experiments showed that all the feature representations learned using cross-lingual BNFs of word pairs outperformed those learning using MFCC features of word pairs. We also investigated the effect of the amount of word-pair supervision on our proposed feature representations. In addition, we performed an invariance test on the learned representations to investigate whether the learned features for each phoneme were relatively more stable with respect to acoustic variations.

2. Methods

We assume that only word pairs are available in a low-resource target language [8, 9, 21]. We use a cross-lingual BNF extractor to parameterize the word pairs of the target language. The cross-lingual BNFs of the word pairs are used for training another NN to generate a new feature representation in the target language.

2.1. Cross-lingual BNFs

We use a BNF extractor which is trained using transcribed data from a high-resource non-target language. The architecture of the BNF extractor is shown in Figure 1. The stacked hierarchical NNs [22] contain two parts: 1) the first-stage NN which takes spectral features as input, and output the first-stage BNFs; 2) the second-stage NN which stacks the first-stage BNFs as input, and output the second-stage BNFs (known as stacked BNFs). Both BNFs extracted from first or second-stage NN work fine, and we choose the latter because they empirically provide better performance in our experiments. Of course, we can simply use the BNF extractor to form a new feature representations for the target language without pairwise supervision

http://dx.doi.org/10.21437/Interspeech.2016-317

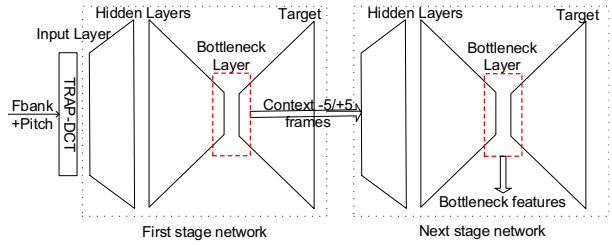

Figure 1: *Cross-lingual BNF extraction.*

of another NN. However, we believe that these cross-lingual BNFs are sub-optimal as their training data are mismatched with target data in languages, so we would obtain an improved feature representation through the pairwise supervision with in-domain data.

2.2. NN representations with word-pair information

The training of NNs with word-pair information as weak supervision has been studied, no matter whether these word pairs are identified by transcriptions [21] or unsupervised term detection [11]. In this paper we consider that it is relatively easy to obtain word pairs identified by native speakers. With word-pair supervision, we classify new feature representations into two types, namely *frame-level* and *word-level* representations. We propose the training procedures for obtaining these two types of new feature representations, and we briefly review some important techniques for learning these two types of new feature representations.

Frame-level feature representations map a sequence of input features to a new sequence of feature vectors with the same length. The procedure for learning our proposed frame-level feature representations are given in Figure 2. First, cross-lingual BNFs are extracted from a cross-lingual NN, and frame alignment between a word pair is done by using dynamic time warping (DTW). Then, we use the aligned frame-pair as input-output with mean squared error (MSE) loss function to train an autoencoder that has been initialized by pretraining. Finally, the last layer of the trained NNs is used to extract the new frame-level feature representations. Since correspondence autoencoders have been shown successful in learning frame-level representations [8], this type of NNs is adopted in our experiments [1].

Word-level feature representations map each whole word segment to a fixed-dimensional vector. The main idea is to find a function which makes same-word pairs more closer and different-word pairs more further. The procedure for learning our proposed word-level feature representations are shown in Figure 3. We use NNs with the Siamese architecture [5] to learn the word-level feature representations. NNs require fixed-dimensional input. If a word segment is shorter than a predefined length (usually defined as the maximum length of all word segments), we pad a word segment into the predefined length by placing the segment in the center and inserting zero vectors into its beginning and end. Since Siamese Convolutional Neural Network (CNN) has been shown successful in learning word-level feature representations for the word discrimination task [9], this type of NNs will be adopted in our experiments. Note that other types of NNs, including time delay neural network (TDNN) [6] and recurrent neural networks (RNN) [24],

[1] Contractive autoencoder [23] has been tried, but with only a small improvement (AP: 0.48).

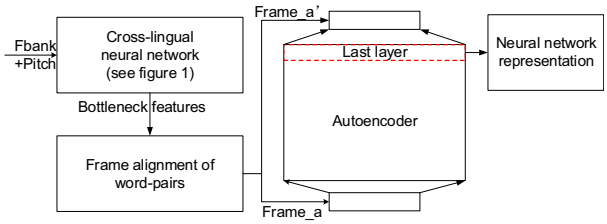

Figure 2: *The scheme of using autoencoder for learning frame-level feature representations.*

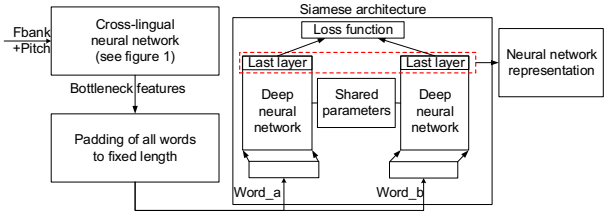

Figure 3: *The scheme of using Siamese network for learning word-level feature representations.*

have been adopted in Siamese architecture for paired supervision in other different tasks. The loss function [2] that we use to train the Siamese architecture is as follows [9, 25]:

$$L = \max\{0, \delta + \frac{1-\cos(x_1, x_2)}{2} - \frac{1-\cos(x_1, x_3)}{2}\} \quad (1)$$

where δ is the margin (set to 0.15 in all of our experiments), x_1 and x_2 are always a pair of same words while x_1 and x_3 are a pair of different words, and x_3 are randomly sampled in each iteration of NN training. In addition, we trained a word classification CNN with word labels, but it did not bring considerable performance gain compared with using weak supervision of word pairs [9].

3. Experiments

3.1. Setup

We evaluated the effectiveness of our proposed feature representations using a word discrimination task on the Switchboard telephone speech corpus. This task calculates the distance between each pair of words and decides whether they are the same or different words. The distance is obtained by performing DTW on the frame-level representations of each word pair, or directly computing the cosine distance of the word-level representations. Average precision (AP), which computes the average value of a precision over the recall interval between 0 and 1, was used to evaluate the performance of each feature representation.

In our experiments, we considered the English speech in the Switchboard corpus as a low-resource target language. We followed the data setup as in [8, 9, 21]. Three non-overlapping sets of 10k, 11k and 11k word tokens (involving around 100 minutes in each set) selected from the corpus were used for feature learning, parameter tuning and test respectively, and each token duration is between 0.5 and 2 seconds. We considered Mandarin Chinese and Spanish as high-resource source languages. We

[2] We also tried different objective functions from [7], but just as explained in [9], the function used in this paper works better.

Table 1: *Average precision (AP) on test set. Pairwise supervision of frame-level (by correspondence autoencoders) and word-level (by Siamese CNNs) feature representations is performed using MFCC features and cross-lingual BNFs of word pairs.*

representations	MFCCs	BNFs (Mandarin)	BNFs (Spanish)
original features	0.214	0.421	0.504
correspondence autoencoder	0.469 [8]	0.619	0.660
Siamese CNN	0.549 [9]	0.678	**0.700**

used around 170 hours of data from the HKUST Mandarin Chinese telephone speech corpus (LDC2005S15) and around 152 hours of data from the Fisher Spanish telephone speech corpus (LDC2010S01) to train the two stacked BNF extractors.

The input features of training the first-stage cross-lingual NNs are 39-dimensional feature vectors, which consist of 36-dimensional Mel filter-bank features and 3-dimensional pitch features. The first-stage cross-lingual NNs used the configuration of 1500-1500-80-1500-X, where the first four numbers indicate the number of neuron units in each layer, and X (equals 412 and 420 in the models for Mandarin Chinese and Spanish respectively) is the number of tied triphone states defined in its initial Gaussian Mixture Model-Hidden Markov Model (GMM-HMM). The input features of training the second-stage cross-lingual NNs were the BNFs, and they were extracted from the first-stage NNs with context expansion that concatenated frames with time offsets [-10,-5,0,+5,+10]. The second-stage NNs also used the configuration of 1500-1500-40-1500-X. No pretraining was used in both two stages of NNs.

Prior to the word-pair supervision of frame-level feature representations, pretraining the stacked autoencoders were performed using 180 conversations (about 23 hours) of speech from the Switchboard telephone speech corpus. The input and output of the correspondence autoencoders are 40-dimensional cross-lingual BNFs. The correspondence autoencoders were constructed by stacking 13 hidden layers, and each hidden layer consisted of 100 units. In addition, the input of a Siamese CNN was two fixed-length sequences of 40-dimensional cross-lingual BNFs from word pair. Each Siamese CNN includes two convolutional and max pooling layers and a fully-connected linear layer with 1024 hidden units. We implemented the correspondence autoencoders and Siamese CNNs based on the open-source code provided by Herman Kamper [3].

3.2. Comparison of different representations in word discrimination task

The performances of different feature representations are shown in Table 1, where each row and column in the table represents an method and its input. The first row of the table shows the performance without any supervision of word pairs. All of the methods based on cross-lingual BNF extractors trained on Mandarin and Spanish significantly outperform those based on MFCC features, including the correspondence autoencoder [8] and Siamese CNN [9] trained on MFCC features of word pairs. This indicates that the information captured in the cross-lingual BNFs for phone classification helps the word discrimination task. Although the structures of cross-lingual BNFs trained on Mandarin Chinese and Spanish are identical, the cross-lingual BNFs trained on Spanish always outperform those trained on Mandarin. We believe that it is because Mandarin Chinese is more different to the target language than Spanish, and the Spanish BNF extractor can capture more useful information for phoneme and word discrimination in the target language. When pairwise supervision on the target language data is performed, the new learned feature representations always outperform their original feature representation. Moreover, Siamese CNN performs better than correspondence autoencoder, no matter which input features we use in the pairwise supervision. The best performance is obtained when the Siamese CNN is trained using the Spanish BNFs of the whole word pairs.

[3]code:https://github.com/kamperh.

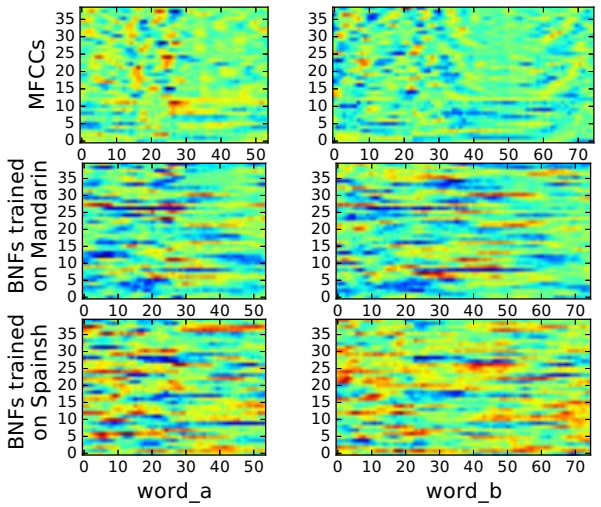

Figure 4: *MFCC features and cross-lingual BNFs of a same-word pair. x-coordinate denotes frames and y-coordinate denotes feature dimensions.*

We choose a word pair from the target data and plot their frame-level features in Figure 4. Horizontal axes denote frames and vertical axes denote feature dimensions. Colors in the plot depict the value of a certain element in a feature vector at a certain frame, with red (blue) indicating large (small) values. More similar and salient horizontal color bands at the same feature dimensions are revealed in the word pair of Mandarin and Spanish BNFs. We believe that these color bands in different feature dimensions form the identity of words or sub-words, which facilitates word discrimination.

3.3. Dependence on the amount of word-pair supervision

To investigate dependence on the amount of word-pair supervision, we varied the number of word pairs N=100k, 10k, 1k, 100 by taking random subsets of the full 100k set as in [8, 9]. Table 2 shows the effect on AP when different numbers of word pairs are used to train correspondence autoencoders. We can find that cross-lingual BNFs based representations consistently outperform those based on MFCC features when different numbers of word pairs are used in frame-level supervision. With 10k word pairs (1/10 of the whole word pairs), cross-lingual BNFs based representations give comparable performance to the whole word pairs. This indicates that cross-lingual BNFs based representations can reach saturation at a faster rate than those MFCC feature based representations. Especially, it is encouraging that the

Table 2: *Average precision (AP) of different frame-level feature representations on test set with different amount of word-pair supervision. N and F indicates the number of word pairs and frame pairs respectively.*

N	F	MFCCs	BNFs (Mandarin)	BNFs (Spanish)
10^5	$7 \cdot 10^6$	0.469	0.619	**0.660**
10^4	$7 \cdot 10^5$	0.385	0.594	**0.660**
10^3	$7 \cdot 10^4$	0.286	0.486	0.554
10^2	$7 \cdot 10^3$	0.259	0.377	0.477

Table 3: *Average precision (AP) of different word-level feature representations on test set with different amount of word-pair supervision. N indicates the number of word pairs.*

N	MFCCs	BNFs (Mandarin)	BNFs (Spanish)
10^5	0.549	0.678	**0.700**
10^4	0.459	0.631	0.656
10^3	0.193	0.386	0.453
10^2	0.067	0.184	0.243

cross-lingual BNFs based representations with no more than 1k word pairs (1/100 of the whole word pairs) can get comparable performance to the previous best results (AP of 0.469) on correspondence autoencoder. This would be practical for the scenario when limited resource is available for annotation of word pairs.

Table 3 shows the effect on AP when different numbers of word pairs are used to train a Siamese CNN. Similarly, cross-lingual BNFs based representations consistently outperform those based on MFCC features when different numbers of word pairs are used in word-level supervision. And the performance is proportional to the amount of word-pair supervision. With 10k word pairs (1/10 of the whole word pairs), the cross-lingual BNFs based representations significantly outperform the previous best performance (AP of 0.549) of those based on MFCC features and trained on all word pairs.

Although Siamese CNNs can outperform correspondence autoencoders when the full set is used, note that they are not as good as correspondence autoencoders when fewer ($<$10k) word pairs are available. It is possibly because Siamese CNN needs more training data to model long temporal information in words.

3.4. Invariance test on the learned feature representations

The learned feature representations have been proved more discriminative in above experiments. To investigate whether they are relatively more stable with respect to acoustic variations, we performed an invariance test on the learned feature representations with respect to each phoneme in the target language. By aligning with word-level transcription to obtain the phoneme label of each frame on the test set, we calculated the average variance of features for each phoneme as follows:

$$average(variance) = \frac{1}{d}\sum_{j=1}^{d}\frac{\sum_{i=1}^{n}(X_{ij}-\overline{X_j})^2}{n-1} \quad (2)$$

where X denotes a matrix which is composed of frame vectors that belongs to the same phoneme, X_j denotes the average value of the j-th column of X (i.e. the average value of the j-th feature dimension), d denotes the feature dimension (i.e. 100), and n denotes the number of frames belonging to this phoneme.

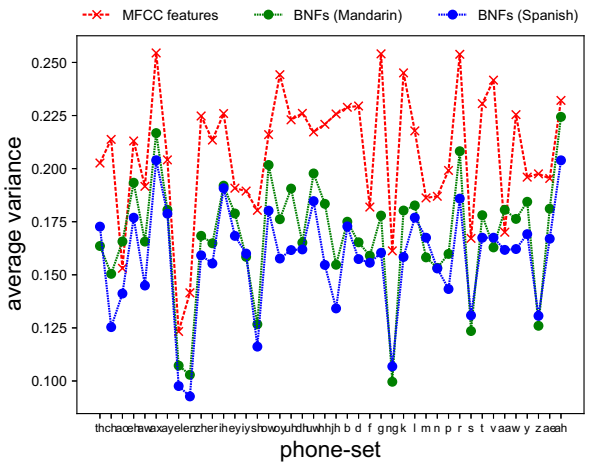

Figure 5: *Phone variance of frame-level learned representations using MFCC features and cross-lingual BNFs.*

The average variance value of each phoneme is shown in Figure 5. It shows that the average variance of each phoneme in cross-lingual BNF based representations is consistently smaller than those in MFCC based representations. Moreover, the average variances in the Spanish BNF based representation are in general smaller than those in the Mandarin Chinese BNF based representation. We believe that our proposed feature representations are relatively less sensitive to acoustic variations (e.g. from environments and speakers) that are common in recordings of human voice.

4. Conclusions

We have proposed a novel way to learn NN representations for a low-resource language. We advocate to use the cross-lingual knowledge obtained from a BNF extractor trained using a high-resource language to parameterize word pairs in the target language. Our proposed method is feasible for learning both frame-level and word-level representations. Our method can provide an average precision (AP) of 0.700, a considerable improvement over the previously best published result on the word discrimination task with a Siamese CNN trained using MFCC features of word pairs. Word-level pairwise learning by the Siamese architecture does not require frame alignment performed by DTW, which is scalable to a large amount of training word pairs. On the other hand, frame-level pairwise learning by correspondence autoencoder is more practical when the budget for obtaining word pairs is limited. Our future work includes: 1) investigating multilingual BNFs [26, 27, 28, 29, 30, 31, 32, 33] for representation learning with word pairs; 2) applying our proposed method to downstream applications such as low-resource ASR and query-by-example spoken term detection.

5. Acknowledgments

We would like to thank Herman Kamper for helpful discussions and the open resource tools, and this work is supported by the National Natural Science Foundation of China (Grant No. 61571363)

6. References

[1] Y. Bengio, "Learning deep architectures for AI," *Foundations and trends® in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[2] G. Hinton, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[3] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, 2015.

[4] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[5] J. Bromley, J. W. Bentz, L. Bottou, I. Guyon, Y. LeCun, C. Moore, E. Säckinger, and R. Shah, "Signature verification using a "Siamese" time delay neural network," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 7, no. 04, pp. 669–688, 1993.

[6] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. CVPR*, 2005, pp. 539–546.

[7] G. Synnaeve, T. Schatz, and E. Dupoux, "Phonetics embedding learning with side information," in *Proc. SLT*, 2014, pp. 106–111.

[8] H. Kamper, M. Elsner, A. Jansen, and S. Goldwater, "Unsupervised neural network based feature extraction using weak top-down constraints," in *Proc. ICASSP*, 2015, pp. 5818–5822.

[9] H. Kamper, W. Wang, and K. Livescu, "Deep convolutional acoustic word embeddings using word-pair side information," in *Proc. ICASSP*, 2016, pp. 4950–4954.

[10] D. Renshaw, H. Kamper, A. Jansen, and S. Goldwater, "A comparison of neural network methods for unsupervised representation learning on the Zero Resource Speech Challenge," in *Proc. INTERSPEECH*, 2015.

[11] A. Jansen and B. Van Durme, "Efficient spoken term discovery using randomized algorithms," in *Proc. ASRU*, 2011, pp. 401–406.

[12] L. Tóth, J. Frankel, G. Gosztolya, and S. King, "Cross-lingual Portability of MLP-Based Tandem Features–A Case Study for English and Hungarian," in *Proc. INTERSPEECH*, 2014, pp. 2695–2698.

[13] P. Swietojanski, A. Ghoshal, and S. Renals, "Unsupervised cross-lingual knowledge transfer in DNN-based LVCSR," in *Proc. SLT*, 2012, pp. 246–251.

[14] A. Stolcke, F. Grezl, M.-Y. Hwang, X. Lei, N. Morgan, and D. Vergyri, "Cross-domain and cross-language portability of acoustic features estimated by multilayer perceptrons," in *Proc. ICASSP*, 2006.

[15] C. Plahl, R. Schlüter, and H. Ney, "Cross-lingual portability of Chinese and English neural network features for French and German LVCSR," in *Proc. ASRU*, 2011, pp. 371–376.

[16] J. Tejedor *et al.*, "Comparison of methods for language-dependent and language-independent query-by-example spoken term detection," *ACM Trans. Inf. Syst.*, vol. 30, no. 3, p. 18, 2012.

[17] Y. Peng *et al.*, "The NNI Query-by-Example System for MediaEval 2014." in *Proc. MediaEval*, 2014.

[18] H. Jingyong *et al.*, "The NNI Query-by-Example System for MediaEval 2015," in *Proc. MediaEval*, 2015.

[19] I. Szoke, M. Skacel, L. Burget, and J. Cernocky, "Coping with channel mismatch in Query-by-Example-But QUESST 2014," in *Proc. ICASSP*, 2015, pp. 5838–5842.

[20] C.-C. Leung *et al.*, "Toward high-performance language-independent query-by-example spoken term detection for MediaEval 2015: post-evaluation analysis," in *Proc. INTERSPEECH*, 2016.

[21] A. Jansen, S. Thomas, and H. Hermansky, "Weak top-down constraints for unsupervised acoustic model training," in *Proc. ICASSP*, 2013, pp. 8091–8095.

[22] F. Grézl and M. Karafiát, "Hierarchical neural net architectures for feature extraction in ASR." in *Proc. INTERSPEECH*, 2010, pp. 1201–1204.

[23] S. Rifai, P. Vincent, X. Muller, X. Glorot, and Y. Bengio, "Contractive auto-encoders: Explicit invariance during feature extraction," in *Proc. ICML*, 2011, pp. 833–840.

[24] J. Mueller and A. Thyagarajan, "Siamese Recurrent Architectures for Learning Sentence Similarity," in *Proc. AAAI*, 2016.

[25] J. Wieting, M. Bansal, K. Gimpel, and K. Livescu, "From paraphrase database to compositional paraphrase model and back," *arXiv preprint arXiv:1506.03487*, 2015.

[26] F. Grézl, M. Karafiát, S. Kontár, and J. Cernocky, "Probabilistic and bottle-neck features for LVCSR of meetings," in *Proc. ICASSP*, 2007, pp. 757–760.

[27] F. Grézl, M. Karafiat, and M. Janda, "Study of probabilistic and bottle-neck features in multilingual environment," in *Proc. ASRU*, 2011, pp. 359–364.

[28] N. T. Vu, F. Metze, and T. Schultz, "Multilingual bottle-neck features and its application for under-resourced languages," in *Proc. SLT*, 2012, pp. 90–93.

[29] K. Veselỳ, M. Karafiát, and F. Grézl, "Convolutive bottleneck network features for LVCSR," in *Proc. ASRU*, 2011, pp. 42–47.

[30] K. Vesely, M. Karafiát, F. Grezl, M. Janda, and E. Egorova, "The language-independent bottleneck features," in *Proc. SLT*, 2012, pp. 336–341.

[31] M. Karafiát, F. Grézl, M. Hannemann, K. Veselỳ, and J. Cernockỳ, "BUT BABEL system for spontaneous Cantonese," in *Proc. INTERSPEECH*, 2013, pp. 2589–2593.

[32] F. Grézl, M. Karafiát, and K. Vesely, "Adaptation of multilingual stacked bottle-neck neural network structure for new language," in *Proc. ICASSP*, 2014, pp. 7654–7658.

[33] Z. Yu, E. Chuangsuwanich, and J. Glass, "Language ID-based training of multilingual stacked bottleneck features," in *Proc. INTERSPEECH*, 2014, pp. 1–5.