A Deep Neural Network Approach for Sentence Boundary Detection in Broadcast News

Chenglin Xu[1,2], *Lei Xie*[1], *Guangpu Huang*[2], *Xiong Xiao*[2], *Eng Siong Chng*[2,3], *Haizhou Li*[2,3,4]

[1]Shaanxi Provincial Key Laboratory of Speech and Image Information Processing, School of Computer Science, Northwestern Polytechnical University, China
[2] Temasek Laboratories@NTU, Nanyang Technological University, Singapore
[3] School of Computer Engineering, Nanyang Technological University, Singapore
[4] Institute for Infocomm Research, A*STAR, Singapore

clxu@mail.nwpu.edu.cn lxie@nwpu.edu.cn gphuang@ntu.edu.sg
xiaoxiong@ntu.edu.sg aseschng@ntu.edu.sg hli@i2r.a-star.edu.sg

Abstract

This paper presents a deep neural network (DNN) approach to sentence boundary detection in broadcast news. We extract prosodic and lexical features at each inter-word position in the transcripts and learn a sequential classifier to label these positions as either boundary or non-boundary. This work is realized by a hybrid DNN-CRF (conditional random field) architecture. The DNN accepts prosodic feature inputs and non-linearly maps them into boundary/non-boundary posterior probability outputs. Subsequently, the posterior probabilities are combined with lexical features and the integrated features are modeled by a linear-chain CRF. The CRF finally labels the inter-word positions as boundary or non-boundary by Viterbi decoding. Experiments show that, as compared with the state-of-the-art DT-CRF approach [1], the proposed DNN-CRF approach achieves 16.7% and 4.1% reduction in NIST boundary detection error in reference and speech recognition transcripts, respectively.

Index Terms: sentence boundary detection, structural event detection, deep neural network, rich transcription

1. Introduction

Adding punctuation makes speech recognition outputs more readable and easier for downstream speech and language tasks, such as parsing, machine translation and question answering. Sentence boundary detection aims to discover sentence boundary positions in an audio stream or in a word transcript provided by a speech recognizer. We usually formulate this task as a binary classification problem which decides if a candidate position, e.g., inter-word in a text or a salient pause in an audio stream, should be marked as a sentence boundary.

In order to train a boundary classifier, previous approaches have explored lexical and prosodic features on both reference transcriptions (REF) and speech recognition outputs (ASR). Nicola et al. [2] studied various lexical features, including language model features, sentence length features and syntax features, on different genres ranging from formal newspaper text to informal, dictated messages, and from written text to spoken transcript. Recent efforts have shown that speech prosody, especially pause and pitch related features, are informative indicators for structural events [1, 3, 4, 5] including sentence boundaries [6, 7, 8, 9]. Research has shown that a decision tree (DT) model learned from prosodic features can achieve comparable performance with that learned from lexical features.

State-of-the-art sentence boundary detection systems usually use features from different knowledge sources. Shriberg et al. [6] integrated both prosodic and lexical features by a decision tree - hidden Markov model (DT-HMM) approach. They first modeled prosodic features using a DT, and the boundary/non-boundary posterior probabilities from the DT were subsequently combined with lexical features in an HMM. Decoding using the HMM results in boundary and non-boundary predictions. The HMM approach has a clear drawback that it maximizes the joint probability of observed and hidden events, as opposed to maximizing the posterior probability that would be a more suitable criterion to the classification task. Recently, conditional random fields (CRFs) have been used in sentence boundary detection and punctuation prediction tasks [1, 10, 11]. As compared with the HMM generative approach, CRF leverages the global sequential information and estimates the posterior probabilities in a discriminative way. Liu et al. [1] proposed a DT-CRF approach. Similar to with the DT-HMM approach, the posterior probabilities from the DT prosodic model were integrated with lexical features in a linear-chain CRF, which led to state-of-the-art sentence boundary detection performance.

In this paper, we present a deep neural network (DNN) approach to sentence boundary detection in broadcast news. In the past several years, DNN and deep learning methods have been successfully used in many tasks, such as speech recognition [12], word segmentation [13, 14], part-of-speech tagging and chunking [15]. A DNN learns a hierarchy of nonlinear feature detectors that can capture complex statistical patterns. Each layer in the DNNs nonlinearly transforms its input representation into a higher level, resulting in a more abstract representation that better models the underlying factors of the data. In our approach, we first model prosodic features using a DNN that accepts prosodic feature inputs and results in boundary/non-boundary predictions with posterior probabilities on the output layer. As compared with the prosodic DT approach [6], a 3-hidden-layer DNN achieves about 11% relative NIST boundary detection error reduction in both REF and ASR broadcast news transcripts. Following the DT-CRF approach, we then integrate the posterior probabilities from the prosodic DNN with the lexical features in a linear-chain CRF, namely the DNN-CRF approach. Experiments show that, as compared with the state-of-the-art DT-CRF approach [1], the proposed DNN-CRF approach achieves 16.7% and 4.1% reduction in NIST boundary

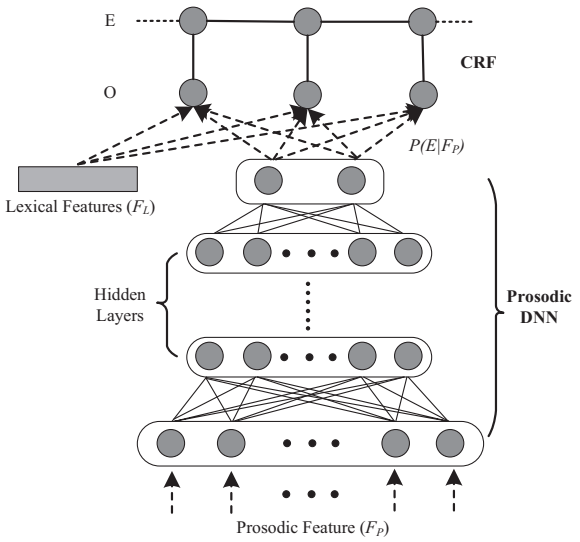

Figure 1: Architecture of the DNN-CRF approach for sentence boundary detection.

detection error in reference and speech recognition transcripts, respectively.

In the following section, we describe our DNN approach for sentence boundary detection. After that, we introduce the prosodic and lexical features in Section 3. Section 4 describes the experiment setup for sentence boundary detection. We report and analyze the experimental results in Section 5. Finally, conclusions are drawn in Section 6.

2. The Proposed Approach

Figure 1 depicts the architecture of the proposed DNN-CRF approach for sentence boundary detection. The architecture is composed of a DNN for prosodic modeling and a linear-chain CRF for sequential boundary/non-boundary labeling based on combined lexical features and DNN posterior probabilities. At each inter-word region across the broadcast news transcript, lexical and prosodic features are extracted and hidden event sequence E is decoded as boundary or non-boundary by the CRF.

A DNN is actually a multi-layer perceptron (MLP), i.e., a feed-forward neural network model that maps sets of input data onto a set of outputs. In our case, the input and output are prosodic features and boundary/non-boundary posterior probabilities, respectively. A DNN can be considered as a hierarchical feature learner with a nonlinear transformation in each hidden layer which refines the input representation to a better one. Each hidden layer takes in the activations $\boldsymbol{h_{l-1}}$ of the previous layer and computes new activations $\boldsymbol{h_l}$ for the next layer via a nonlinear transformation using a weight matrix $\boldsymbol{W_l}$ and a bias vector $\boldsymbol{b_l}$ followed by an activation function $f_l(\cdot)$:

$$\boldsymbol{h_l} = f_l(\boldsymbol{W_l h_{l-1}} + \boldsymbol{b_l}), \text{ for } 1 \leq l \leq L. \tag{1}$$

Where L is the number of hidden layers. In this paper, we use sigmoid activation functions for hidden layers. The output layer adopts a softmax function to predict the posterior probabilities for each of the classes (boundary and non-boundary) given the input observation (prosodic features) using a weight matrix $\boldsymbol{W_L}$ and a bias vector $\boldsymbol{b_L}$.

The posterior probabilities $P(E|F_P)$ are further combined with lexical features F_L and the combined feature sequence O are modeled by a linear-chain CRF. The CRF integrates different knowledge sources in a discriminative way and leverages the sequential and contextual information. A CRF defines a conditional probability distribution $P(E|O)$ of corresponding label sequence E given input observation sequence O [16]. In this work, E corresponds to a boundary/non-boundary event sequence. The most likely label sequence $\hat{E}$ for given observation O is:

$$\begin{aligned} \hat{E} &= \arg\max_E P(E|O) \\ &= \arg\max_E \frac{\exp(\sum_k^K \lambda_k * F_k(E,O))}{\sum_E \exp(\sum_k^K \lambda_k * F_k(E,O))} \end{aligned} \tag{2}$$

where $F_k(E,O)$ is a feature function over the labels and observations. The index k indicates different feature function, each of which has an associated weight λ_k. For an input sequence O and a label sequence E, $F_k(E,O)$ is defined as:

$$F_k(E,O) = \sum_i f_k(E,O,i) \tag{3}$$

where i is the index over all the input positions. $f_k(E,O,i)$ is the feature function at position i over the label sequence and observation sequence.

The CRF model assigns a well-defined conditional probability distribution over possible labels on a given training set, trained by the maximum likelihood criterion. Its loss function is convex that guarantees convergence to the global optimum. The Viterbi algorithm is used to find the most likely label sequence.

When $f_k(E,O,i) = f_k(E_{i-N},...,E_i,O_{i-M},...,O_i,i)$, an N-order linear-chain CRF, which models N $(E = E_{i-N},...,E_i)$ sequence labels and M $(O = O_{i-M},...,O_i)$ context features in the feature set, is formed. In practice, $N = 1$ and $M = 1$ are usually used because of the exponential increase of computational cost for higher N and M.

3. Feature Extraction

3.1. Prosodic Features

Speech prosodic cues, e.g., pitch, energy and duration, are known to convey structural information. Previous studies have shown that they play important roles in boundary perception [17, 18, 19]. In our study, we consider the inter-word positions across a broadcast news transcript as boundary candidates and collect a rich set of 162 prosodic features in the audio stream corresponding to the candidate positions according to the method in [6, 20]. Please refer to [6, 20] for feature extraction details. Among the features, pause and word duration features are used to capture prosodic continuity and boundary lengthening phenomena. We also extract pitch and energy related features that reflect the pitch/energy declination and reset phenomena. These features have been shown as primary cues for sentence boundary detection [1, 6, 8]. In broadcast news, as speaker turn is a significant boundary cue, we also include speaker turn as a feature.

3.2. Lexical Features

According to [1], we extract lexical features that include N-grams of word, part-of-speech (POS) tag and syntactic chunk tag. It is well-known that the lexical context of sentence boundary is important for boundary detection. In order to capture the

word context of sentence boundary, we use word N-grams (up to 5) features, i.e., $< w_i >$, $< w_{i-1}, w_i >$, $< w_i, w_{i+1} >$, $< w_{i-2}, w_{i-1}, w_i >$, $< w_{i-1}, w_i, w_{i+1} >$, $< w_i, w_{i+1}, w_{i+2} >$ and $< w_{i-2}, w_{i-1}, w_i, w_{i+1}, w_{i+2} >$, where w_i refers to the word before the boundary of interest. Each sentence is constrained via syntactic structure. Therefore, syntactic tags (e.g., POS and Chunk) constitute a prominent knowledge source for sentence boundary detection. In this paper, we use the SENNA parser [15] to obtain the POS sequence (p) and the chunk sequence (c) given a word stream. The IOBES tagging scheme is used for chunking so as to map the word sequence to chunk stream exactly like POS. POS and chunk features are designed similar to those in words, replacing w_i with tags p_i and c_i.

4. Experimental Setup

We evaluate the performance of sentence boundary detection using our proposed approach on English broadcast news (BN). The BN data comes from NIST RT-04F and RT-03F MDE evaluation [1]. The released corpora from LDC only contain the training set of the evaluations (about 40 hours). In order to keep our experimental configuration (hours of training data) as similar as possible to [1], we extract 2-hour data from the RT-04F released data as the testing set. Another 2-hour data is selected as the development set for parameter tuning. The rest of the data (36 hours) is used as the training set. Meanwhile, we repeat the state-of-the-art method in [1] with our experimental configuration as a comparision. The sentence boundaries in reference transcripts (REF) are annotated according to the annotation guideline [21]. The recognition outputs (ASR) are generated from an in-house speech recognizer with a word error rate of 29.5%. In the data, about 8% of the inter-word positions are sentence boundaries.

For the sentence boundary detection task, we train all the models using REF transcripts, and evaluate the models on both REF and ASR transcripts. Evaluation across REF and ASR transcripts allows us to study the influence of speech recognition errors. Evaluation metrics include precision, recall, F1-measure and the NIST SU error rate. The SU error rate is defined as the total number of inserted and deleted boundaries divided by the number of real boundaries. We calculate SU error using the official NIST evaluation tools[2].

The prosodic DNN is trained in a greedy layer-wise supervised training way [22, 23]. We start with 1-hidden layer neural network that maps prosodic features into boundary/non-boundary posterior probabilities. After the network is trained, treat the output of the hidden layer as new features and train another 1-hidden layer network to predict the boundary/non-boundary posteriors. The procedure can repeat until the desirable number of hidden layers are reached and finally, a fine tuning of the whole network is performed. The training is implemented by using stochastic gradient descent (SGD) and the minibatch size is 256 shuffled training samples. As our train data size is small, to prevent overfitting, $L2$ weight decay is set to 0.00001. Furthermore, the system development data is split into training data and validation data. Network training is stopped once the error on the validation data starts to increase.

We compare the DNN-CRF approach with the DT-CRF approach [1] that obtains state-of-the-art performance. A C4.5 decision tree is built using the WEKA toolkit [3] based on the prosodic features. The DT posteriors are combined with the lexical features by a CRF sequential labeler. For our DNN-CRF approach, similarly, we use a CRF to combine the DNN posteriors with the lexical features. The CRF++ toolkit is used for CRF implementation [4]. Because the toolkit can only handle discrete features, we follow [1] and quantize the posterior probabilities into several bins: $[0, 0.1]$, $(0.1, 0.3]$, $(0.3, 0.5]$, $(0.5, 0.7]$, $(0.7, 0.9]$, $(0.9, 1]$.

5. Results and Discussion

5.1. Results of Prosodic DNN

In this section, we evaluate the performance of the DNN model in sentence boundary detection only using prosodic features with different hidden layers and number of hidden units. For clarity, we only use NIST SU error rate as the evaluation criterion. Figure 2a shows the effects of using different numbers of hidden layers in a DNN. We can see that, on the REF transcripts, the NIST SU error rate obtained by DNN is much lower than that obtained by DT. In addition, the SU error rate decreases with the increase of network depth until 3 hidden layers. The result of DNN with 3 hidden layers and different number of hidden units is drawn in Figure 2b. Best performance is obtained when the number of hidden units is set to 80. In summary, the best DNN setting is 3 hidden layers each with 80 nodes and we will always use this setting in the following experiments.

Table 1: Experimental comparison of the prosodic DNN and prosodic DT approach. Results are reported using Precision (P), Recall (R), F1-measure (F1) and NIST SU error rate (NIST).

Transcript	Approach	P / R / F1 (%)	NIST (%)
REF	DT	78.8 / 56.3 / 65.7	58.8
	DNN	86.9 / 56.5 / 68.5	**52.1**
ASR	DT	70.6 / 56.7 / 62.9	67.0
	DNN	74.3 / 61.7 / 67.4	**59.7**

Table 1 summarizes the results of DT and DNN in both REF and ASR test conditions. From the table, we can observe that the prosodic DNN significantly outperforms the prosodic DT in both REF and ASR conditions (significant at $p < 0.05$ [24] for SU error rate). As compared with the prosodic DT approach, DNN achieves 11.4% and 10.9% relative NIST SU error reduction for REF and ASR conditions, respectively. The performance gain is mainly attributed to the DNN's ability to learn prominent representations from a large raw feature set through several non-linear transform stages. We also notice an increase of SU error rate for both DT and DNN on ASR transcriptions. This is mainly because the word errors in recognition outputs affect the prosodic feature extraction. For example, the wrong word timing information misleads the prosody extraction region, since we choose the inter-word boundary as the candidates. However, we observe that DT suffers more from the recognition errors than DNN. This may indicate that DNN is more robust in processing the imperfect prosodic features.

[1] LDC2005S16, LDC2004S08 for speech data and LDC2005T24, LDC2004T12 for reference transcriptions

[2] See http://www.itl.nist.gov/iad/894.01/tests/rt/2004-fall/

[3] Available at: http://www.cs.waikato.ac.nz/ml/weka/index.html

[4] Available at: https://code.google.com/p/crfpp/

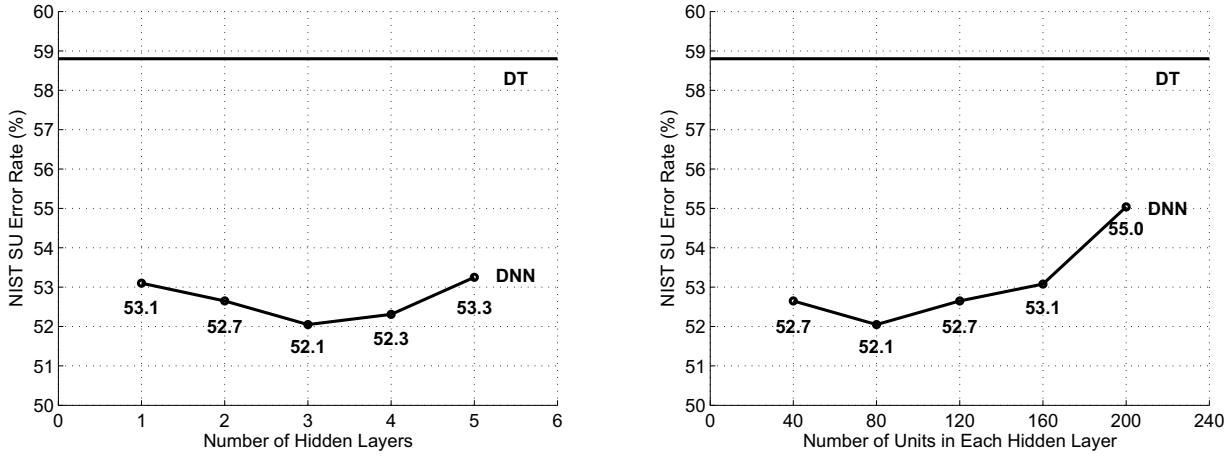

Figure 2: Effects of DNNs with different hidden layers and units. (a) Effects of different hidden layers in a DNN. The units in each layer are kept as 80. (b) Effects of a DNN with different units in each hidden layer. The DNN has 3 hidden layers.

Table 2: Experimental comparison between DT-CRF and DNN-CRF in REF and ASR conditions.

Approach	Information Source		REF		ASR	
	Lexical	Prosodic	P / R / F1 (%)	NIST (%)	P / R / F1 (%)	NIST (%)
DT-CRF [1]	Word-POS-Chunk	DT Posterior	81.4 / 73.9 / 77.4	43.1	90.6 / 49.5 / 64.0	55.6
DNN-CRF		DNN Posterior	85.9 / 76.7 / 81.0	**35.9**	95.0 / 49.3 / 64.9	**53.3**

5.2. Results of DNN-CRF

Table 2 shows the performances of DT-CRF and DNN-CRF, both combine the lexical features with prosodic posterior probabilities. The results show that combining lexical and prosodic information generally results in better performance. We believe the significant performance gain comes from two aspects. First, lexical features, especially POS and Chunk features, are very helpful in sentence boundary detection because the POS and Chunk information reflects the syntactic structure of a sentence. Second, CRF effectively leverages the sequential information in sentence boundary detection. There is one case where adding lexical features leads to worse results, i.e., the recall of sentence boundaries is reduced to around 49% when ASR transcripts are used. At the same time, precision is increased dramatically to above 90%. From these two results, we can conclude that the using of imperfect ASR transcripts leads to significantly more missing sentence boundaries. Word recognition errors may mislead the POS and Chunk tagging, and the prosody model is also affected since the prosodic features are extracted with imperfect word transcripts.

Another observation is that DNN-CRF always outperforms DT-CRF. However, the improvement on REF transcripts (from 43.1% to 35.9% SU error rate, i.e., 16.7% relative reduction) is much larger than the improvement on ASR transcripts (from 55.6% to 53.3%, i.e., 4.1% relative reduction). Although the improvements are both significant ($p < 0.01$ for REF, and $p < 0.05$ for ASR), the results show that we have less gain from DNN prosodic model when ASR transcripts are used. Possible reason could be the high word error rate (29.5%) of our ASR system. Comparing the precision and recall obtained from ASR transcripts, DNN-CRF obtained better precision (95.0%) than DT-CRF (90.6%). However, as the recall is very low for both systems, the final F1 measure is not improved much. In the future, we will focus on improving the recall of the DNN-CRF system when using ASR transcripts with high word error rate.

6. Conclusion

We have proposed a deep neural network approach for sentence boundary detection in broadcast news. In our approach, we first use a DNN to model prosodic features extracted at each inter-word positions in the broadcast news transcripts. The prosodic DNN achieves significant performance gain as compared with the DT approach. Subsequently, we use a CRF to combine the posterior probabilities from the prosodic DNN with extracted lexical features. The CRF finally labels the inter-word positions as boundaries or non-boundaries. Experiments show that the proposed DNN-CRF approach outperforms the state-of-the-art DT-CRF approach [1] by a large margin. Future work goes in two directions. First, as DNN has shown its superior performance in multi-task training [25], we plan to explore its ability in multilingual sentence boundary detection. Second, we plan to test different neural networks, e.g., convolution neural networks (CNN) [26] and recurrent neural networks (RNN) [27], in sentence boundary detection.

7. Acknowledgements

This work was supported by a grant from the National Natural Science Foundation of China (61175018) and a grant from the Fok Ying Tung Education Foundation (131059).

8. References

[1] Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. Harper, "Enriching speech recognition with automatic detec-

tion of sentence boundaries and disfluencies," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1526–1540, 2006.

[2] U. Nicola, B. Maximilian, and V. Paul, "Improved models for automatic punctuation prediction for spoken and written text," in *Proceedings of INTERSPEECH*, 2013.

[3] L. Xie, "Discovering salient prosodic cues and their interactions for automatic story segmentation in mandarin broadcast news," *Multimedia Systems*, vol. 14, no. 4, pp. 237–253, 2008.

[4] F. Batista, D. Caseiro, N. Mamede, and I. Trancoso, "Recovering capitalization and punctuation marks for automatic speech recognition: Case study for portuguese broadcast news," *Speech Communication*, vol. 50, no. 10, pp. 847–862, 2008.

[5] X. Wang, L. Xie, M. Lu, B. Ma, E. S. Chng, and H. Li, "Broadcast news story segmentation using conditional random fields and multimodal features," *IEICE TRANSACTIONS on Information and Systems*, vol. 95, no. 5, pp. 1206–1215, 2012.

[6] E. Shriberg, A. Stolcke, D. Hakkani-Tür, and G. Tür, "Prosody-based automatic segmentation of speech into sentences and topics," *Speech communication*, vol. 32, no. 1, pp. 127–154, 2000.

[7] Y. Liu, A. Stolcke, E. Shriberg, and M. Harper, "Using conditional random fields for sentence boundary detection in speech," in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL)*. Association for Computational Linguistics, 2005, pp. 451–458.

[8] L. Xie, C. Xu, and X. Wang, "Prosody-based sentence boundary detection in chinese broadcast news," in *Proceedings of the 8th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2012, pp. 261–265.

[9] C. Xu, L. Xie, and F. Zhonghua, "Sentence boundary detection in Chinese broadcast news using conditional random fields and prosodic features," in *Proceedings of ChinaSIP*. IEEE, 2014.

[10] X. Wang, H. Ng, and K. Sim, "Dynamic conditional random fields for joint sentence boundary and punctuation prediction," in *Proceedings of Interspeech*, 2012.

[11] W. Lu and H. Ng, "Better punctuation prediction with dynamic conditional random fields," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2010, pp. 177–186.

[12] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[13] X. Zheng, H. Chen, and T. Xu, "Deep learning for Chinese word segmentation and POS tagging," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2013, pp. 647–657.

[14] K. Evang, V. Basile, G. Chrupała, and J. Bos, "Elephant: Sequence labeling for word and sentence segmentation," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2013.

[15] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *The Journal of Machine Learning Research*, vol. 12, pp. 2493–2537, 2011.

[16] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," 2001.

[17] C. Tseng, S. Pin, Y. Lee, H. Wang, and Y. Chen, "Fluent speech prosody: Framework and modeling," *Speech Communication*, vol. 46, no. 3, pp. 284–309, 2005.

[18] Y. Mo, "Duration and intensity as perceptual cues for naïve listeners prominence and boundary perception," in *Proceedings of the 4th Speech Prosody Conference*, 2008, pp. 739–742.

[19] T. Mahrt, J. Cole, M. M. Fleck, and M. Hasegawa-Johnson, "F0 and the perception of prominence." in *Proceedings of INTERSPEECH*, 2012.

[20] Z. Huang, L. Chen, and M. Harper, "An open source prosodic feature extraction tool," in *Proceedings of the Language Resources and Evaluation Conference (LREC)*, 2006.

[21] S. Strassel, "Simple metadata annotation specification v6.2," in *http://www.ldc.upenn.edu/Projects/MDE*. LDC, 2004.

[22] Y. Bengio, P. Lamblin, D. Popovici, H. Larochelle *et al.*, "Greedy layer-wise training of deep networks," *Advances in neural information processing systems*, vol. 19, p. 153, 2007.

[23] H. Larochelle, Y. Bengio, J. Louradour, and P. Lamblin, "Exploring strategies for training deep neural networks," *The Journal of Machine Learning Research*, vol. 10, pp. 1–40, 2009.

[24] P. Koehn, "Statistical significance tests for machine translation evaluation." in *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2004, pp. 388–395.

[25] L. Deng, J. Li, J.-T. Huang, K. Yao, D. Yu, F. Seide, M. Seltzer, G. Zweig, X. He, J. Williams *et al.*, "Recent advances in deep learning for speech research at microsoft," in *Proceedings of 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2013, pp. 8604–8608.

[26] O. Abdel-Hamid, L. Deng, and D. Yu, "Exploring convolutional neural network structures and optimization techniques for speech recognition," in *Proceedings of Interspeech*, 2013.

[27] A. Graves, A. R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proceedings of 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2013, pp. 6645–6649.