An End-to-End Neural Network Approach to Story Segmentation

Jia Yu*, Lei Xie*‡ , Xiong Xiao†, Eng Siong Chng†,
* Shaanxi Provincial Key Laboratory of Speech and Image Information Processing,
School of Computer Science, Northwestern Polytechnical University, Xi'an, China
† School of Computer Engineering, Nanyang Technological University, Singapore
E-mail: {jiayu,lxie}@nwpu-aslp.org, {xiaoxiong,ASESChng}@ntu.edu.sg

***Abstract*— This paper proposes an end-to-end story segmentation approach based on long short-term memory (LSTM) - recurrent neural network (RNN). Traditional story segmentation approaches are a two-stage pipeline consisting of feature extraction and segmentation, each of which has its individual objective function. In other words, the objective function used to extract features is different from the true performance measure of story segmentation, which may degrade the segmentation results. In this paper, we combine the two components and optimize them jointly, using an LSTM-RNN. Specifically, one LSTM layer is used to extract sentence vectors, and another LSTM layer is used to predict story boundaries by taking as input of the sentence vectors. Importantly, the whole network is optimized directly to predict story boundaries. We also investigate bi-directional LSTM (BLSTM) that can utilize past and future information in the process of extracting sentence vectors and story boundary prediction. Experimental results on the TDT2 corpus show that the proposed approach achieves state-of-the-art performance in story segmentation.**

I. Introduction

Story segmentation is a task of partitioning a stream of audio, video or text into story segments, each addressing a specific topic. It is a necessary precursor for a variety of language processing technologies including content indexing and retrieval [1], document summarization [2], topic detection and tracking [3], [4] and information extraction [5]. Typical story segmentation approaches are a pipeline consisting of feature learning and segmentation. The two components are not optimized jointly for story segmentation, making independent assumptions for individual components [6], [7], [8], [9]. Recently, end-to-end (E2E) neural network (NN) learning that jointly optimizes all components (e.g., in speech recognition) has achieved promising results [10], [11], [12]. This motivates us to develop an end-to-end NN approach for the story segmentation task at hand.

Story segmentation has been studied for different genres, such as broadcast news [13], [14], meeting recordings [15] and lectures [16], [17], etc., over various types of media, including audio [17], [18], [19], video [20] and text [21], [22], [23], [24], [6], [15]. In this paper, we aim to perform story segmentation for textual documents like broadcast news speech recognition transcripts. Note that, with the recent tremendous success of large vocabulary continuous speech recognition (LVCSR) using deep neural networks (DNN) [25], [26], [27], [28], [29], [30], [31], we can easily obtain high accuracy transcripts for broadcast news. Thus traditional text segmentation approaches, with similar purposes of story segmentation, can be easily applied to the speech recognition transcripts.

Traditional story segmentation approaches on texts are a pipeline system consisting of feature learning that catches semantic or topic information from a stream of text, and segmentation that partitions the stream to topically coherent segments by detecting the topic shift.

Feature extraction heavily affects the performance of story segmentation. Bag-of-words (BOW) representation, or term frequency-inverse document frequency (tf-idf), is a simple representation in typical story segmentation approaches, e.g., TextTiling and dynamic programming (DP) [6], [7], [8]. However, BOW or tf-idf only counts the appearances of words, ignoring semantic relations among them. Instead, probabilistic latent semantic analysis (pLSA) [9], latent Dirichlet allocation (LDA) [32], and LapPLSA [33], employ latent topic variables and create topic model that depicts the probability distribution of words on topics. With these probabilistic models, BOW based word representations are transformed into topic representations and used in varies segmentation approaches [32], [34]. Recently, neural network based topic models have shown promising performances [35], [36], [37], [38], [39]. Specifically, we derived word representation in topic space from a neural network based topic model, leading to improved story segmentation performance [40].

The second component of the pipeline is a segmenter. The above-mentioned TextTiling [6], [7] and dynamic programming (DP) [33], [41], [42], [43] are typical detection-based approaches, which find optimal partitions over word sequence by optimizing a local or global objective. Popular probabilistic model approaches locate story boundaries by probability distribution of topics on document and probability distribution of words on topics. Popular such approaches include PLSA [34], BayesSeg [44], dd-CRP [45] and HMM [23], [24], [21].

The two components of a story segmentation system are traditionally modeled independently. The objective function used to extract feature may be substantially different from the true performance measure of story segmentation. This sort of inconsistency may degrade the performance of story segmentation. The purpose of end-to-end (E2E) learning is to

‡Corresponding author

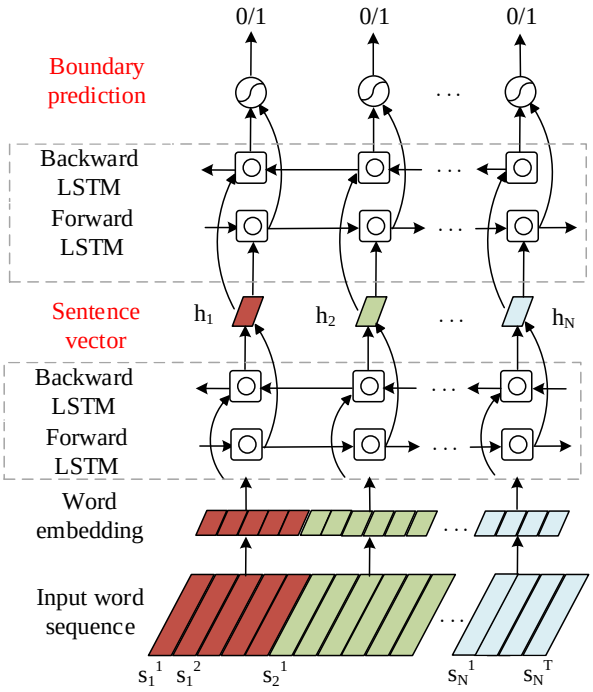

Fig. 1. Diagram of the proposed end-to-end story segmentation approach.

combine different components in the pipeline as a whole and optimize them jointly. There are several major advantages for end-to-end learning:

- The whole model is more closely related to the target since it has one objective function.
- It is more efficient because large computational flow graphs can be optimized together by simple back propagation in the training process.
- The whole system is quite simple since there are only one input and one output and features are automatically learned in the end-to-end network.

Recently, E2E learning has achieved promising results in various tasks. In speaker identification, an E2E i-vector system was built by combining speaker (i-vector) feature extraction and identification together using a deep neural network architecture [46], [47], [48]. An E2E speech recognition system [10] was proposed to replace various components, e.g., pronunciation dictionary and input-output alignments, in the traditional pipeline. This was achieved by a deep bidirectional long short-term memory (LSTM) recurrent neural network (RNN) and a connectionist temporal classification (CTC) objective function. With a trigram language model, the system has achieved comparable performance to traditional approaches. Deep Speech, an end-to-end speech recognition system trained using well-optimized RNN and multiple GPUs, has achieved 16% error on Switchboard Hub5'00 corpus and can handle challenging noisy environments [49]. Deep Speech 2 was proposed to recognize either English or Mandarin Chinese speech by using an end-to-end learning strategy that can handle a diverse variety of speech including noisy environments, accents and different languages [50].

Inspired by the success of E2E systems in the forementioned tasks, we propose an end-to-end learning approach for story segmentation using LSTM-RNN. Specifically, we use one LSTM layer for feature extraction, which forms sentence vectors by accumulating word sequence information. The derived sentence vectors are fed into another LSTM layer that captures the context information of each sentence. Finally, a sigmoid layer is used to predict story boundaries. Importantly, the whole network is optimized under a unique cost function for story boundary prediction. We also investigated bi-directional LSTM (BLSTM) layers as an alternative, since they can accumulate both past and future information [51]. Experimental results on the TDT2 corpus show that the proposed approach achieves state-of-the-art performance in story segmentation.

II. The Proposed Approach

A. LSTM for Sentence Embedding and Story Boundary Prediction

LSTM has strong capability of capturing long term information and hence we utilize it to extract fixed dimensional sentence vectors and predict boundary information. As shown in Fig. 1, the input of the neural network is a sequence of sentences: $\boldsymbol{s} = [\boldsymbol{s}_1, \boldsymbol{s}_2, ..., \boldsymbol{s}_N]$, where N denotes the number of sentences in the input text stream. Each sentence is comprised of T words in the form of 1-hot representation, and T varies from sentence to sentence. The words of 1-hot representation in the n-th sentence, represented by $\boldsymbol{s}_n = [\boldsymbol{s}_n^1, \boldsymbol{s}_n^2, ..., \boldsymbol{s}_n^T]$, are projected into word vectors by a linear projection layer and then fed into the LSTM layer one-by-one. The LSTM layer goes through the sentence, increasingly accumulates the information, and the output of last word is regarded as the representation of whole sentence. Thus the sequence of sentences $[\boldsymbol{s}_1, \boldsymbol{s}_2, ..., \boldsymbol{s}_N]$ is transformed to sequence of sentence vectors $[\boldsymbol{h}_1, \boldsymbol{h}_2, ..., \boldsymbol{h}_N]$, and then fed into another LSTM layer. For each sentence $\boldsymbol{h}_i$, the LSTM layer captures its context information, followed by an output layer that predicts whether it is at a position of story boundary.

Different from the conventional pipeline systems in which each component is modeled separately, in the proposed approach, training the whole network is achieved by simply error back propagation (BP) [52] on a training set of texts, in which each sentence is labeled with a boundary identification (0/1) based on whether it is at the position of a story boundary. Specially, the training process includes forward propagation and backward propagation. The forward propagation calculates the prediction errors and the backward propagation reversely passes the errors back to modify the model parameters.

B. Bidirectional LSTM

One shortcoming of LSTM is that it only makes use of past information. However, in the process of extracting sentence-level features and boundary prediction, past and future context are obviously both important. Thus we try to use bi-

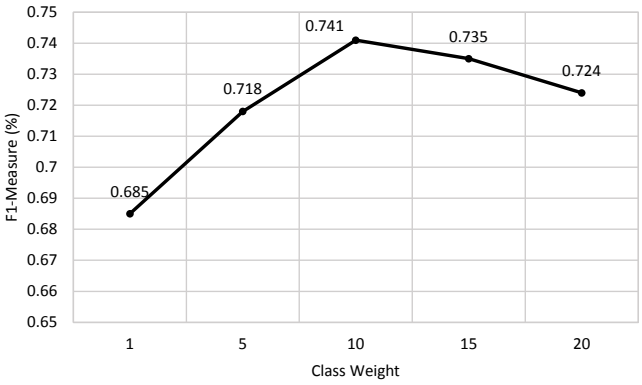

Fig. 2. F1-measure with different class weight.

directional LSTM (BLSTM) that captures the data flow from both directions. BLSTM consists of a forward LSTM and a backward LSTM, and the outputs of the two sub-networks are finally combined. Specifically, given an input sequence, the forward LSTM reads it from left to right and the backward LSTM reads it in a reversed order. Each LSTM has their own parameters. In this way, the BLSTM can use both past inputs and future inputs for a specific time. For convenience, we name the proposed end-to-end story segmentation approach with LSTM and BLSTM layers as E2E-LSTM and E2E-BLSTM, respectively.

III. Experiments

A. Experimental Setup

We carried out experiments on the Topic Detection and Tracking (TDT2) corpus [4] which includes 2,280 English broadcast news programs. There were 11,406 stories in total and each story had an average of 20 topics and 200 words. We constructed a vocabulary including 57,817 words. The corpus was separated into a training set with 1,800 programs, a development set and a test set each with 240 programs. All texts were stemmed by a Porter stemmer and stop words were removed. As sentence delimiters are not available in transcriptions, according to [53], we used significant pauses as delimiters to construct pseudo-sentences. We used Keras toolkit [54] to build neural networks. The input data has 3 dimension shapes which are block, sentence and word. Each block contains the same number of sentences that is 5 in this study, while each sentence contains varies number of words. As Keras only accepts sequences of the same length in a batch, we padded input sentences to the same length of 20 words that is close to the average number of words in a sentence in the training data. Hierarchical LSTM architecture is used in the neural network, specifically, we used one LSTM layer to construct sentence vectors from a sequence of words, and another LSTM layer to predict whether the current sentence is a boundary given its context information which depends on the length of the block.

We used F1-measure, i.e., the harmonic mean of recall and precision, to evaluate the story segmentation performance with a tolerance window of 50 words according to the TDT2 standard [4]. The discovered boundaries were compared to the manually segmented boundaries. Precision is defined as the percentage of declared boundaries that coincide with the referenced boundaries. Recall is defined as the percentage of referenced boundaries that are retrieved. Thus F1-measure is defined as

$$\text{F1-measure} = 2 \times \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}. \quad (1)$$

We treated the story boundary detection problem as a sequential labeling task, with sentence boundaries being identified as either a story boundary or not at each inter-sentence position in the input text stream. In text documents like broadcast news transcripts, non-boundary sentences are far more than those boundary sentences, which causes data imbalance problem. There are two possible ways to handle such a situation: over or down sampling and using a weighted error measure during training [55], [56], [57]. The former methods have the problem of inconsistency of class distribution between training and test sets, which will affect the classification performance. Thus we use the latter strategy in which the errors are computed in proportion to the class weight.

TABLE I
F1-measure with different number of LSTM layers and LSTM nodes

LSTM layers# \ LSTM nodes#	256	512	768	1024
1	0.741	0.757	0.752	0.736
2	0.759	**0.772**	0.764	0.748
3	0.74	0.755	0.749	0.735

B. Results of E2E-LSTM

Fig. 2 shows the segmentation performance with different values of class weight. The x-axis is the weight ratio of boundary and non-boundary, which is denoted as $1 : x$, where x ranges from 1 to 20. Y-axis denotes the value of F1-measure. The ratio of $1 : 1$ means we assign equal class weight to the loss function for the two categories. We used one LSTM layer with 256 nodes for sentence vector extraction and another LSTM layer with same nodes for boundary prediction. From the figure, we can observe that the value of F1-measure boosted drastically as x increases from 1 to 10, and reaches the highest value when the ratio is 1:10. Then the F1-measure begins to decrease when x surpasses 10. This observation demonstrates that assigning an appropriate weight to the minor class for the loss function can significantly improve the performance for the imbalanced data problem at hand.

Table I shows the story segmentation performance with different numbers of LSTM layers and nodes for sentence vector extraction and boundary prediction. The number of LSTM layers and nodes ranges from 1 to 3 and 256 to 1024, respectively. The ratio of class weight is fixed to 1:10 in the experiment. According to the results, we can conclude that the F1-measure is significantly improved when the number of

TABLE II
F1 WITH DIFFERENT NUMBER OF BLSTM LAYERS AND NODES

BLSTM layers# \ BLSTM nodes#	256	512	768	1024
1	0.749	0.762	0.755	0.747
2	0.766	**0.779**	0.771	0.759
3	0.735	0.758	0.752	0.741

TABLE III
THE PERFORMANCE OF E2E-LSTM AND E2E-BLSTM SYSTEMS ARE BOTH IMPROVED WHEN WE ADD MORE TRAINING DATA. THE TEST SET IS NOT CHANGED.

Network Type	E2E-LSTM	E2E-BLSTM
F1-measure	0.776	0.787

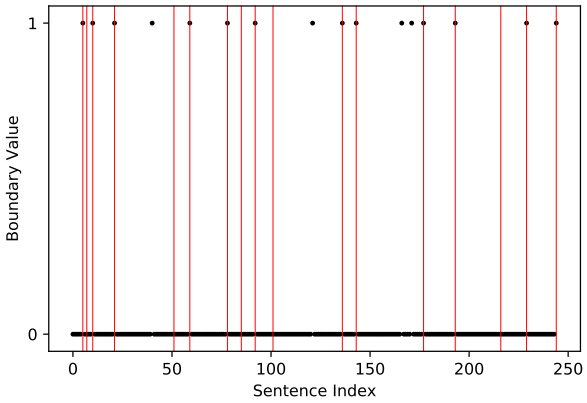

Fig. 3. Boundaries predicted by the proposed E2E-BLSTM approach for an episode of broadcast news program from TDT2 corpus. X-axis is index of sentences, and 0 and 1 on y-axis represent non-boundary and boundary, respectively. The red vertical lines indicate real topic boundaries.

TABLE IV
F1-MEASURE COMPARISON WITH STATE-OF-THE-ART METHODS.

Approach	F1-measure
TextTiling [6]	0.553
HMM [23]	0.637
PLSA-DP-CE [33]	0.682
BayesSeg [44]	0.710
DD-CRP [45]	0.730
DNN-HMM [58]	0.765
LSTM-HMM [59]	0.774
E2E-LSTM (this study)	0.772
E2E-BLSTM (this study)	0.779
E2E-LSTM (with extra TDT4 data)	0.776
E2E-BLSTM (with extra TDT4 data)	**0.787**

LSTM layers is increased to 2, and then begins to decrease when increased to 3. With a fixed number of LSTM layers, we obtain the best performance when the number of LSTM nodes is 512, as shown by the rows in Table I. We obtain the highest F1-measure of 0.772 from a neural network with 2 LSTM layers and 512 nodes, used for both sentence embedding and boundary prediction.

C. Results of E2E-BLSTM

We further tested the performance of the proposed approach with different BLSTM layers. The results are summarized in Table II. Compared with Table I, we notice that the F1-measure is improved with the same neural network architecture when the LSTM layers are replaced by the BLSTM layers. When the neural network has 2 BLSTM layers and 512 nodes, used for both sentence embedding and boundary prediction, we obtain the highest F1-measure of 0.779. The better performance achieved by E2E-BLSTM indicates that bidirectional contexts are quite useful for story segmentation.

As the proposed E2E model is to directly learn boundaries from text streams, it is easy to recruit more training data from other data sets. Thus we included extra training data from the TDT4 corpus which has another 1,400 broadcast news programs. The performance of E2E-LSTM and E2E-BLSTM were both improved. The F1-measure was improved to 0.776 and 0.787 for E2E-LSTM and E2E-BLSTM, respectively, compared to 0.772 and 0.779 with only the TDT2 corpus as the training data. Please note that the results were obtained from the same testing set.

We also visualized the predicted story boundaries and compared them with the true boundaries. Specifically, Fig. 3 shows the boundaries predicted by E2E-BLSTM for an episode of broadcast news program from the TDT2 corpus. The x-axis is the index of sentences, and 0 and 1 on the y-axis represent non-boundary and boundary, respectively. A sentence with a value of 1 on y-axis indicates that it is at the position of a story boundary. The red vertical lines indicate real story boundaries. From the figure we can see, the predicted boundaries follow the true boundaries reasonably well, which indicates the effectiveness of the proposed E2E learning approach.

D. Comparison with the state-of-the-art methods

Finally, we compared the proposed approach with some state-of-the-art methods benchmarked on the TDT2 corpus. The results are summarized in Table IV. We can clearly see the proposed E2E-BLSTM approach outperforms all the methods in the comparison. The F1-measure of E2E-LSTM is slightly lower than LSTM-HMM approach (0.772 and 0.774, respectively). However, when we use BLSTM layers, the performance surpasses that of the LSTM-HMM approach and achieves the best performance.

IV. CONCLUSIONS

In this paper, we have proposed an end-to-end neural network approach to story segmentation. Different from conventional approaches that treat feature extraction and boundary prediction separately, we use a neural network to model and optimize the two components jointly. Specifically, in our neural network, we use an LSTM layer to extract sentence-level embedding and another LSTM layer to predict story boundaries. Importantly, the whole network is optimized under a unique cost function for story boundary prediction. We also investigated using BLSTM layers that can capture both past and future information. Experiments on TDT2 corpus show that the proposed approach outperforms the traditional ap-

proaches and achieves state-of-the-art performance. In future, we plan to study pair-wise learning [60], [61], [62] in end-to-end story segmentation.

V. Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 61571363).

References

[1] L.-s. Lee and B. Chen, "Spoken document understanding and organization," *Signal Processing Magazine, IEEE*, vol. 22, no. 5, pp. 42–60, 2005.

[2] L. F. Rau, P. S. Jacobs, and U. Zernik, "Information extraction and text summarization using linguistic knowledge acquisition," *Information Processing & Management*, vol. 25, no. 4, pp. 419–428, 1989.

[3] A. James, "Introduction to topic detection and tracking," *Topic detection and tracking*, pp. 1–16, 2002.

[4] J. Fiscus, G. Doddington, J. Garofolo, and A. Martin, "Nists 1998 topic detection and tracking evaluation (tdt2)," *Proceedings of the 1999 DARPA Broadcast News Workshop*, pp. 19–24, 1999.

[5] S. Soderland, "Learning information extraction rules for semi-structured and free text," *Machine learning*, vol. 34, no. 1-3, pp. 233–272, 1999.

[6] M. A. Hearst, "Texttiling: Segmenting text into multi-paragraph subtopic passages," *Computational linguistics*, vol. 23, no. 1, pp. 33–64, 1997.

[7] L. Xie, Y.-L. Yang, and Z.-Q. Liu, "On the effectiveness of subwords for lexical cohesion based story segmentation of chinese broadcast news," *Information Sciences*, vol. 181, no. 13, pp. 2873–2891, 2011.

[8] A. Bouchekif, G. Damnati, and D. Charlet, "Intra-content term weighting for topic segmentation," *in Proc. ICASSP*, pp. 7113–7117, 2014.

[9] M. Lu, C.-C. Leung, L. Xie, B. Ma, and H. Li, "Probabilistic latent semantic analysis for broadcast news story segmentation." in *Proc. INTERSPEECH*, 2011, pp. 1109–1112.

[10] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *Proc. ICML*, 2014, pp. 1764–1772.

[11] S. Kim, T. Hori, and S. Watanabe, "Joint ctc-attention based end-to-end speech recognition using multi-task learning," *CoRR*, vol. abs/1609.06773, 2016.

[12] Y. Zhang, W. Chan, and N. Jaitly, "Very deep convolutional networks for end-to-end speech recognition," *CoRR*, vol. abs/1610.03022, 2016.

[13] A. Rosenberg and J. Hirschberg, "Story segmentation of brodcast news in english, mandarin and arabic," in *Proc. HLT*, 2006, pp. 125–128.

[14] H. Chen, B. Guo, Z. Yu, and Q. Han, "Toward real-time and cooperative mobile visual sensing and sharing," in *Proc. INFOCOM*, 2016, pp. 1–9.

[15] S. Banerjee and A. I. Rudnicky, "A texttiling based approach to topic boundary detection in meetings," in *Proc. ICSLP*, 2006.

[16] I. Malioutov and R. Barzilay, "Minimum cut model for spoken lecture segmentation," in *Proc. ACL*, 1998, pp. 25–32.

[17] I. Malioutov, A. Parkand, R. Barzilay, and J. Glass, "Making sense of sound: Unsupervised topic segmentation over acoustic input," in *Proc. ACM*, 2007, p. 504.

[18] E. Shriberg, A. Stolcke, D. Hakkani-Tur, and G. Tür, "Prosody-based automatic segmentation of speech into sentences and topics," *Speech Communication*, vol. 32, no. 1-2, pp. 127–154, 2000.

[19] D. Charlet, G. Damnati, A. Bouchekif, and A. Douib, "Fusion of speaker and lexical information for topic segmentation: A co-segmentation approach," in *Proc. ICASSP*, 2015, pp. 5261–5265.

[20] L. Chaisorn, T. S. Chua, and C. H. Lee, "A multi-modal approach to story segmentation for news video," *World Wide Web-internet and Web Information Systems*, vol. 6, no. 2, pp. 187–208, 2003.

[21] J. P. Yamron, I. Carp, L. Gillick, S. Lowe, and P. van Mulbregt, "A hidden markov model approach to text segmentation and event tracking," in *Proc. ICASSP*, 1998, pp. 333–336.

[22] D. Beeferman, A. Berger, and J. Lafferty, "Statistical models for text segmentation," *Machine learning*, vol. 34, no. 1-3, pp. 177–210, 1999.

[23] M. Sherman and Y. Liu, "Using hidden markov models for topic segmentation of meeting transcripts," in *Proc. SLT*, 2008, pp. 185–188.

[24] P. Van Mulbregt, I. Carp, L. Gillick, S. Lowe, and J. Yamron, "Text segmentation and topic tracking on broadcast news via a hidden markov model approach." in *Proc. ICSLP*, 1998.

[25] D. Yu and L. Deng, *Automatic Speech Recognition - A Deep Learning Approach*. Springer, 2015.

[26] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. ICASSP*, 2013, pp. 6645–6649.

[27] O. Abdel-Hamid, L. Deng, and D. Yu, "Exploring convolutional neural network structures and optimization techniques for speech recognition." in *Proc. INTERSPEECH*, 2013, pp. 3366–3370.

[28] D. Yu, M. L. Seltzer, J. Li, J.-T. Huang, and F. Seide, "Feature learning in deep neural networks-studies on speech recognition tasks," *arXiv preprint arXiv:1301.3605*, 2013.

[29] T. Schultz and A. Waibel, "Language-independent and language-adaptive acoustic modeling for speech recognition," *Speech Communication*, vol. 35, no. 1, pp. 31–51, 2001.

[30] B. Damavandi, S. Kumar, N. Shazeer, and A. Bruguier, "Nn-grams: Unifying neural network and n-gram language models for speech recognition," in *Proc. INTERSPEECH*.

[31] H. A. Bourlard and N. Morgan, *Connectionist speech recognition: a hybrid approach*. Springer Science & Business Media, 2012, vol. 247.

[32] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[33] M. Lu, L. Zheng, C.-C. Leung, L. Xie, B. Ma, and H. Li, "Broadcast news story segmentation using probabilistic latent semantic analysis and laplacian eigenmaps," *in Proc. APSIPA ASC 2011*, pp. 356–360, 2011.

[34] Hofmann and Thomas, "Probabilistic latent semantic indexing," in *Proc. SIGIR*, 1999, pp. 50–57.

[35] L. Wan, L. Zhu, and R. Fergus, "A hybrid neural network-latent topic model." in *Proc. AISTATS*, vol. 12, 2012, pp. 1287–1294.

[36] C. Li, H. Wang, Z. Zhang, A. Sun, and Z. Ma, "Topic modeling for short texts with auxiliary word embeddings," in *Proc. SIGIR*, 2016, pp. 165–174.

[37] S. Lai, L. Xu, K. Liu, and J. Zhao, "Recurrent convolutional neural networks for text classification," *In: Proc. national conferece on artificial intelligence*, pp. 2267–2273, 2015.

[38] H. Larochelle and S. Lauly, "A neural autoregressive topic model," *Advances in Neural Information Processing Systems*, pp. 2708–2716, 2012.

[39] G. Kumar and L. F.D'Haro, "Deep autoencoder topic model for short texts," in *Proc. IWES*, 2015.

[40] J. Yu, X. Xiao, L. Xie, E. S. Chng, and H. Li, "A DNN-HMM approach to story segmentation," in *Proc. INTERSPEECH*, 2016, pp. 1527–1531.

[41] P. Fragkou, V. Petridis, and A. Kehagias, "A dynamic programming algorithm for linear text segmentation," *Journal of Intelligent Information Systems*, vol. 23, no. 2, pp. 179–197, 2004.

[42] O. Heinonen, "Optimal multi-paragraph text segmentation by dynamic programming," in *Proc. ACL*, 1998, pp. 1484–1486.

[43] L. Xie, L. Zheng, Z. Liu, and Y. Zhang, "Laplacian eigenmaps for automatic story segmentation of broadcast news," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 1, pp. 276–289, 2012.

[44] J. Eisenstein and R. Barzilay, "Bayesian unsupervised topic segmentation," in *Proc. EMNLP*, 2008, pp. 334–343.

[45] C. Yang, L. Xie, and X. Zhou, "Unsupervised broadcast news story segmentation using distance dependent chinese restaurant processes," in *Proc. ICASSP*, 2014, pp. 4062–4066.

[46] C. Li, X. Ma, B. Jiang, X. Li, X. Zhang, X. Liu, Y. Cao, A. Kannan, and Z. Zhu, "Deep speaker: an end-to-end neural speaker embedding system," *CoRR*, vol. abs/1705.02304, 2017.

[47] S. Zhang, Z. Chen, Y. Zhao, J. Li, and Y. Gong, "End-to-end attention based text-dependent speaker verification," *CoRR*, vol. abs/1701.00562, 2017.

[48] D. Snyder, P. Ghahremani, D. Povey, D. Garcia-Romero, Y. Carmiel, and S. Khudanpur, "Deep neural network-based speaker embeddings for end-to-end speaker verification," in *Proc. SLT*, 2016, pp. 165–170.

[49] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates *et al.*, "Deep speech: Scaling up end-to-end speech recognition," *arXiv preprint arXiv:1412.5567*, 2014.

[50] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *Proc. ICML*, 2016, pp. 173–182.

[51] D. Amodei, R. Anubhai, E. Battenberg, and C. Case, "Deep speech 2 : End-to-end speech recognition in english and mandarin," in *Proc. ICML*, 2016, pp. 173–182.

[52] J. Li, J. H. Cheng, J. Y. Shi, and F. Huang, *Advances in Computer Science and Information Engineering*. Springer Berlin Heidelberg, 2012.
[53] H. Chen, L. Xie, W. Feng, L. Zheng, and Y. Zhang, "Topic segmentation on spoken documents using self-validated acoustic cuts," *Soft Computing-A Fusion of Foundations, Methodologies and Applications*, vol. 19, no. 1, pp. 47–59, 2015.
[54] F. Chollet, "Keras (2015)," *URL http://keras. io*, 2017.
[55] M. A. Mazurowski, P. A. Habas, J. M. Zurada, J. Y. Lo, J. A. Baker, and G. D. Tourassi, "Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance," *Neural Networks*, vol. 21, no. 2-3, pp. 427–436, 2008.
[56] S. H. Khan, M. Bennamoun, F. A. Sohel, and R. Togneri, "Cost sensitive learning of deep feature representations from imbalanced data," *CoRR*, vol. abs/1508.03422, 2015.
[57] F. J. Pulgar, A. J. Rivera, F. Charte, and M. J. del Jesús, "On the impact of imbalanced data in convolutional neural networks performance," in *Proc. HAIS*, 2017, pp. 220–232.
[58] J. Yu, X. Xiao, L. Xie, E. S. Chng, and H. Li, "A dnn-hmm approach to story segmentation," in *Proc. INTERSPEECH*, 2016, pp. 1527–1531.
[59] J. Yu, L. Xie, X. Xiao, and E. S. Chng, "A hybrid neural network hidden markov model approach for automatic story segmentation," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–12, 2017.
[60] J. Mueller and A. Thyagarajan, "Siamese recurrent architectures for learning sentence similarity." in *AAAI*, 2016, pp. 2786–2792.
[61] Y. Yuan, C.-C. Leung, L. Xie, B. Ma, and H. Li, "Learning neural network representations using cross-lingual bottleneck features with word-pair information." in *Proc. INTERSPEECH*, 2016, pp. 788–792.
[62] Y. Yuan, C.-C. Leung, L. Xie, H. Chen, B. Ma, and H. Li, "Pairwise learning using multi-lingual bottleneck features for low-resource query-by-example spoken term detection," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 5645–5649.