Hierarchical recurrent neural network for story segmentation using fusion of lexical and acoustic features

HIERARCHICAL RECURRENT NEURAL NETWORK FOR STORY SEGMENTATION USING FUSION OF LEXICAL AND ACOUSTIC FEATURES

Emiru Tsunoo[1,2]*, Ondřej Klejch*[1]*, Peter Bell*[1]*, Steve Renals*[1]

[1]Center for Speech Technology Research, School of Informatics
University of Edinburgh, Edinburgh EH8 9AB, UK
[2]System R&D Group, R&D Platform
Sony Corporation, Tokyo 141-8610, Japan

ABSTRACT

A broadcast news stream consists of a number of stories and it is an important task to find the boundaries of stories automatically in news analysis. We capture the topic structure using a hierarchical model based on a Recurrent Neural Network (RNN) sentence modeling layer and a bidirectional Long Short-Term Memory (LSTM) topic modeling layer, with a fusion of acoustic and lexical features. Both features are accumulated with RNNs and trained jointly within the model to be fused at the sentence level. We conduct experiments on the topic detection and tracking (TDT4) task comparing combinations of two modalities trained with limited amount of parallel data. Further we utilize additional sufficient text data for training to polish our model. Experimental results indicate that the hierarchical RNN topic modeling takes advantage of the fusion scheme, especially with additional text training data, with a higher F1-measure compared to conventional state-of-the-art methods.

Index Terms— spoken document processing, recurrent neural network, topic modeling, story segmentation, multimodal features

1. INTRODUCTION

The aim of story segmentation is to divide a sequential stream of text or audio into stories or topics. It is useful for many subsequent tasks such as summarization, topic detection, and information retrieval, and plays a crucial role for analyzing media streams. In this paper we are concerned with the segmentation of broadcast media using a combination of acoustic and lexical features, based on a hierarchical model in which each story is assumed to consist of several sentences in a coherent order, and each sentence consists of words which are assumed to be relevant to the story.

Story segmentation has been studied for decades, through various media types such as text [1, 2, 3, 4, 5, 6, 7], audio [8, 9, 10], and video [11, 12, 13]. The studies using text were pioneered by the TextTiling approach [2], where adjacent sentence blocks were compared using a similarity measure based on bag-of-words (BOW) features, such as term frequency - inverted document frequency (*tf-idf*). Later studies indicated that globally optimized segmentation methods – such as dynamic programming (DP) and the hidden Markov model (HMM) [3, 4, 14] – can improve the performance, and usage of probabilistic topic modeling such as probabilistic latent semantic analysis (pLSA) [15, 7] and latent Dirichlet allocation (LDA) [16, 17] can further increase the accuracy. Analogous to approaches used in automatic speech recognition (ASR), deep neural networks have been combined with HMMs (DNN-HMM) and successfully applied to story segmentation, using BOW features of text data, with significant improvement in performance [18]. DNNs have been also applied to similar applications including dialogue session segmentation [19] and sentence boundary detection or punctuation estimation [20, 21]. On the other hand, the studies using acoustic features include Shriberg et al. [8] where pause, phone/rhyme duration, F0 contours and its quality indicators are used, and Rosenberg et al. [10] where statistics of F0 and speaking rate are used. Similar features are also utilized with vision features which are tailored for TRECVID project [12, 13].

Recurrent neural networks (RNNs) have made a great impact on language modeling. Following the feed-forward neural prediction language model [22], Mikolov et al. proposed using an RNN for language modelling, thus removing the limitation of finite context for predicting next words [23]. Language modelling using long short-term memory (LSTM) RNNs was proposed [24], and currently represents the state-of-the-art in language modelling [25]. To incorporate additional context, the paragraph embedding vector was introduced as an auxiliary input to an RNN language model [26, 27], and was found to improve the quality of modeling. This model factorizes into a topic factor and a word distribution

for the topic, with the paragraph vector being trained to represent the topic. Hierarchical models have also been proposed for topic/document modeling [28, 29], and Lin et al. [30] extended the paragraph vector language model using a hierarchical RNN. In this work a sentence-level RNN was used to convey an unlimited history of sentences, and by using this history vector in a similar way to a paragraph vector, each word was predicted with a word-level RNN. We have previously proposed a hierarchical RNN model which is a reverse form of Lin's model, where each sentence is represented as a sentence embedding vector with a word-level RNN layer and overall story transition is modeled with bidirectional LSTM layer, applying the model successfully to story segmentation [31].

In this paper, we extend our hierarchical RNN model to use a fusion of acoustic and lexical features, and apply it to story segmentation. Acoustic features are accumulated into a vector representation and concatenated at the sentence level using a lexical sentence embedding computed with a word-level RNN layer. The overall story transition is modeled with a bidirectional LSTM layer using this fused representation, and finally a feed-forward neural network layer predicts the topic label of the input sentence, followed by an HMM decoder which predicts story boundaries. We also address the realistic scenario in which news audio/text parallel training data is limited, while additional text data is sufficiently available from other news sources. Our model is trained using the parallel topic detection and tracking dataset (TDT4) and additional text from the TDT2 dataset. The model is evaluated on TDT4 with human transcriptions and also on ASR transcriptions, and compared to the state-of-the-art DNN-HMM story segmentation method [18].

2. STORY SEGMENTATION WITH RECURRENT NEURAL NETWORK

2.1. General Formulation of Story Segmentation

Broadcast news consists of various topics and the story segmentation task is to find boundaries between the topics. By considering topics as hidden states, the Hidden Markov Model is widely used for this task [3, 18, 32]. We assume that sentence boundaries are available, similar to [18], as many studies regarding sentence segmentation and punctuation estimation have been done [20, 21]. Given a sequence of sentences $\mathbf{s} = [s_1, ..., s_J]$ and the parameter set θ, we optimize to find the most probable topic label sequence $\hat{\mathbf{z}}$, considering all possible sequences of topic labels $\mathbf{z} = [z_1, ..., z_J]$.

$$\hat{\mathbf{z}} = \underset{\mathbf{z}}{\operatorname{argmax}}\ p(\mathbf{z}|\mathbf{s};\theta) \tag{1}$$

Analogous to a DNN-HMM acoustic model, this optimization problem can be solved with a combination of topic

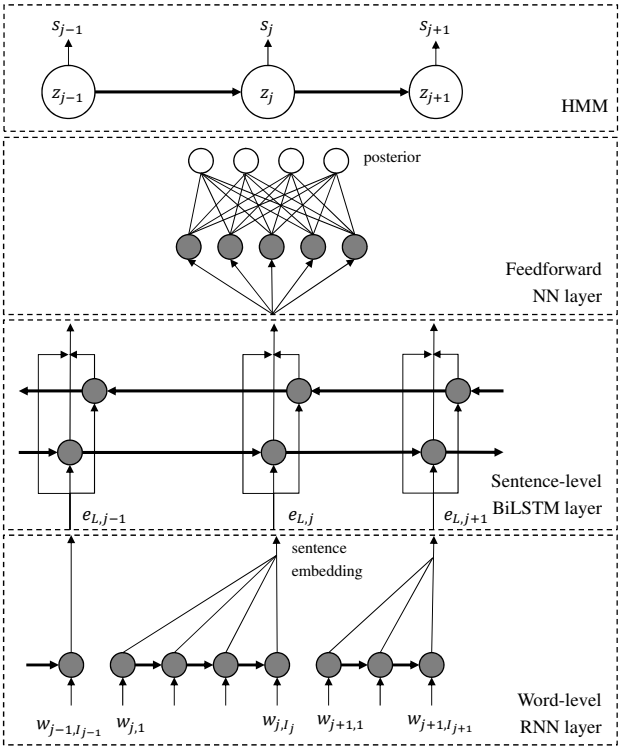

Fig. 1. Hierarchical recurrent neural network for story segmentation.

posterior prediction, $p(z_j|s_j)$, and transition probability modeling, $p(\mathbf{z})$, by applying Bayes' rule:

$$\begin{aligned}\hat{\mathbf{z}} &= \underset{\mathbf{z}}{\operatorname{argmax}}\ p(\mathbf{s}|\mathbf{z};\theta)p(\mathbf{z})/p(\mathbf{s}) \\ &= \underset{\mathbf{z}}{\operatorname{argmax}}\ p(\mathbf{s}|\mathbf{z};\theta)p(\mathbf{z})\end{aligned} \tag{2}$$

$$p(s_j|z_j) = \frac{p(z_j|s_j)}{p(z_j)}p(s_j). \tag{3}$$

$p(\mathbf{s})$ and $p(s_j)$ do not depend to $\mathbf{z}$ and can be ignored. $p(z_j)$ is considered as prior probability, and the topic posterior $p(z_j|s_j)$ can be estimated using the proposed hierarchical RNN. The prior probability of the sequence $p(\mathbf{z})$ is modelled via HMM transition probabilities.

2.2. Hierarchical Recurrent Neural Network with Lexical Features

Broadcast news has a hierarchical character, with a top level sequence of stories, in which each story consists of multiple sentences, and each sentence consists of words which are relevant to the story. To capture this structure, we have proposed a hierarchical RNN model combining a sentence embedding RNN and a bidirectional LSTM story transition model [31],

depicted in Figure 1. In the first layer, the word-level sentence embedding RNN, independently concentrates each sentence into a sentence embedding vector. This is followed by the second layer which models the transition of multiple stories within a chunk, for instance a program unit, using a sentence-level bidirectional RNN which considers a context of both preceding and following sentences. The final feed-forward layer estimates topic posterior probabilities which may be used in an HMM to decode the topic sequence – as in Equation (3) – thus obtaining the story boundaries.

We utilize a bypass technique [31] which allows the model to use not only the outputs of the bidirectional LSTM layer but also the output of the RNN sentence embedding layer directly. Let the sentence embedding vector $e_{L,j}$ for sentence j be defined as

$$e_{L,j} = \sum_{i=1}^{I_j} \lambda_{j,i} h_{j,i} \tag{4}$$

where L indicates a lexical embedding, I_j is the total number of words in the sentence j, $\lambda_{j,i}$ are predefined weights, and $h_{j,i}$ is the history vector of the word-level RNN given the i-th word embedding vector $w_{j,i}$ as input. The weight parameters $\lambda_{j,i}$ can be all set to 0 except for last word which is set to 1 to filter out only the last history vector (cf. [33]). They can be also set equally to $1/I_j$ so that the gradients spread to every time step in order to avoid the problem of vanishing or exploding gradients. In addition, let the output vectors of both sentence-level forward and backward LSTM be $h_{F,j}$ and $h_{B,j}$. Then the posterior $p(z_j|s_j)$ is calculated as following with the last feed-forward neural network layer,

$$y_j = \sigma(W_F h_{F,j} + W_B h_{B,j} + W_r e_{L,j} + b_y) \tag{5}$$

$$p(z_j|s_j) = g(W_p y_j + b_p) \tag{6}$$

where σ is sigmoid function, g represents softmax function, and matrices W_* and bias vectors b_* are trainable.

2.3. Fusion Scheme with Acoustic Features

2.3.1. Acoustic feature extraction

Acoustic prosody can convey additional information about story boundaries. For instance, prosodic changes may coincide with story change points, hence a variety of prosodic features have been used to detect story boundaries [8, 10, 12, 13]. In this paper, we propose to utilize prosody features in addition to log-filterbank energy features. We use normalized voicing intensity and pitch as the prosody features based on the autocorrelation calculation used in YIN [34]. If t is the frame number in sentence j then the normalized voicing intensity $v_{j,t}$ and pitch $l_{j,t}$ are calculated as

$$v_{j,t} = \max_{\tau} \frac{\tau d_{j,t}(\tau)}{\sum_{n=1}^{\tau} d_{j,t}(n)} \tag{7}$$

$$l_{j,t} = \arg\max_{\tau} \frac{\tau d_{j,t}(\tau)}{\sum_{n=1}^{\tau} d_{j,t}(n)} \tag{8}$$

where

$$d_{j,t}(\tau) = \sum_{n=1}^{W} (x_{j,n} - x_{j,n+\tau})^2, \tag{9}$$

in which $x_{j,n}$ is an input signal and W is the analysis window size. By concatenating these prosody features with filterbank features, we obtain acoustic features $a_{j,t}$.

2.3.2. Fusion with statistical features

Summary statistics of acoustic features are widely used for speech segmentation [8, 10, 12, 13]. Hsu et al. reported that pause duration and pitch jump were the dominant acoustic features in their maximum entropy multi-modal model [12]. Therefore, following [13], we extract the mean, variance, minimum, and maximum of the pause durations, voiced segments, and pitch from the acoustic features $a_{j,t}$, from 1 second after the previous sentence and for the entire current sentence. A pause is defined as a region which continuously satisfies $v_{j,t} < \delta$ where δ is typically set to 0.5 since $v_{j,t}$ has a range of $[0, 1]$, and vice versa for the voiced segments. We also extract the pitch jump with respect to the previous sentence. In addition, we also use the mean and variance of the filterbank features and their delta coefficients. The extracted statistical features are concatenated with lexical sentence embedding as in Figure 2-(a). Ideally, the fusion model is trained with parallel data of audio and text.

2.3.3. Fusion with sentence embedding

An acoustic embedding can also be computed from the acoustic features $a_{j,t}$ for sentence j using an RNN, similar to the lexical embedding described in Section 2.2. The lexical and acoustic embeddings may be fused (Fig. 2-(b)), and the accumulated information passed to the upper bidirectional LSTM layer. The RNN is trained jointly with the hierarchical model. The acoustic embedding $e_{A,j}$ is calculated in a similar way to (4), using the history vector $h'_{j,t}$ of RNN with given acoustic features $a_{j,t}$, as

$$e_{A,j} = \sum_{t=1}^{T_j} \lambda'_{j,t} h'_{j,t} \tag{10}$$

where A represents an acoustic embedding, T_j is a total number of frames in sentence j and $\lambda'_{j,t}$ are weight parameters.

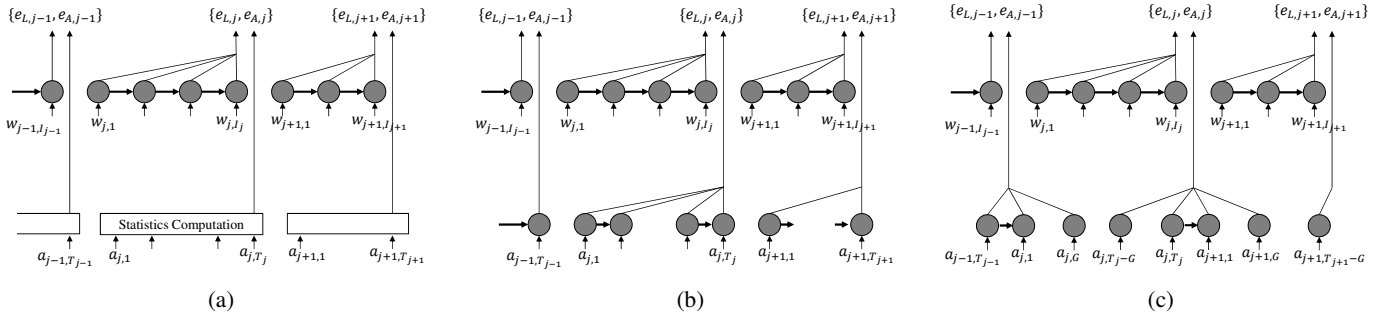

Fig. 2. Feature fusion in three different ways: (a) - Concatenation of acoustic feature statistics. (b) - Concatenation of acoustic sentence representation using RNN. (c) - Concatenation of acoustic sentence gap representation using RNN.

2.3.4. Fusion with sentence gap embedding

Instead of embedding complete sentences, it is possible to consider only the information around sentence boundaries. Therefore we also propose an acoustic embedding based on a fixed time window around the sentence boundaries, using the final G frames of acoustic features in the current sentence and the following G frames, which are fed into the acoustic embedding RNN. We refer to this embedding as sentence gap embedding, and it is calculated similar to (10):

$$e_{A,j} = \sum_{t=T_{j-1}-G}^{T_{j-1}} \lambda'_{j-1,t} h'_{j-1,t} + \sum_{t=1}^{G} \lambda'_{j,t} h'_{j,t} \qquad (11)$$

This fusion is depicted in Figure 2-(c).

2.4. Training Procedure

The training was done by minimizing the cross-entropy between the target probabilities and the output posterior $p(z_j|s_j)$ using gradient descent, with the target probabilities provided according to predefined cluster labels. However, in general, it is not easy to obtain data which has topic labels, and therefore, the labels were estimated by unsupervised clustering using CLUTO [35], similar to [18]. Based on a *tf-idf* representation, topic segments were clustered by minimizing the inter-cluster similarity and maximizing the intra-cluster similarity, each sentence within a segment being labeled according to the clusters.

To align the input lexical tokens and audio signals, the acoustic features were extracted from the audio data using a 10 ms frame rate, and we used ASR models to align the acoustic features and lexical tokens obtained from a human transcription. The sequence of acoustic features is divided into sentences at the point where each sentence ends, i. e., any pause or sound after each sentence end is included in the following sentence chunk.

Table 1. Dataset Specification.

Dataset	TDT4 [37]		TDT2 [38]	
	Text (# of tokens)	Audio (minutes)	Text (# of tokens)	Audio (minutes)
Training	364,218	5,345	7,004,119	–
Validation	39,738	618	1,013,940	–
Test	156,630	2,299	–	–

In order to generalize the training, the broadcast program units were broken into story segments, shuffled, and concatenated again into pseudo-programs of average program size. In that manner we created as many possible combination of stories as possible. The word-level lexical RNNs and acoustic RNNs were duplicated by the number of sentences in a program unit and connected in parallel to a sentence-level LSTM. The parameters were initialized with random values ranging from -0.1 to 0.1 except bias vectors, which were set to 0, and updated for every pseudo program unit. The gradients for first RNN layer were clipped if their norm exceeded 0.5 to avoid the exploding gradients problem [36]. The learning rate α was set to 1 at the beginning and changed to $\alpha/2$ if the loss for validation set increased. The training process was terminated after 30–40 epochs.

3. EXPERIMENTS

3.1. Experimental Setup

We evaluated our hierarchical RNN on the Topic Detection and Tracking (TDT4) task [37]. For testing, a randomly chosen set of 78 programs out of the TDT4 data were used. All words in the data were preprocessed by the Porter stemmer and stop words were removed. The specification for testing data is shown in "Test" row in Table 1.

The lexical word-level RNN, sentence-level bidirectional

Table 2. F1-measure comparison of combinations of features using models trained with TDT4 parallel data.

Model	F1-Measure
Lexical DNN [18]	0.711
Lexical HRNN [31]	0.689
Fusion HRNN (statistics)	**0.742**
Fusion HRNN (sentence embedding)	0.738
Fusion HRNN (sentence gap embedding)	0.724
Fusion DNN (statistics)	0.706

Table 3. F1-measure comparison of combinations of features using models trained with TDT4 parallel data and additional TDT2 text data.

Model	F1-Measure
Lexical DNN [18]	0.718
Lexical HRNN [31]	**0.738**
Fusion HRNN (statistics)	0.729
Fusion HRNN (sentence embedding)	**0.755**
Fusion HRNN (sentence gap embedding)	0.750
Fusion DNN (statistics)	0.726

LSTM and feed-forward neural network all used 256 hidden units, and the word embedding input vector was also trained using 256 dimensions. For acoustic features, 40 log-filterbank features were computed from 0–4000 Hz and prosodic features were calculated using (7) and (8) with $W = 280$; the acoustic RNN used 32 hidden units. For sentence gap embedding, we applied the RNN to the final 1 second of the end of each sentence and the following 1 second in the next sentence, i.e., $G = 100$ in (11). The weight parameters $\lambda_{j,i}$ and $\lambda'_{j,t}$ for both lexical and acoustic RNN sentence embedding were set to the uniform values, $1/I_j$ and $1/T_j$ respectively, taking the average of the history vectors. For each HMM state, the transition probability of staying in the same state was set to 0.8, with the remaining transition probability evenly divided between the other states [18, 32].

Story boundaries were detected as change points of the topic sequence decoded by the HMM, and evaluated using the F1-measure[1] comparing with the manual segment boundary annotation. Our method was tested using 150 clusters which was found to be optimal in our previous work [31]. We compared with the state-of-the-art method, DNN-HMM story segmentation [18].

3.2. Evaluations on Acoustic and Lexical Fusion

First, we trained our model using only the TDT4 data, with a training set of 180 programs and a test set of 20 programs. Training and test data statistics are given in Table 1.

The results of the conventional DNN model and the proposed hierarchical RNN model with only lexical features are shown as Lexical DNN and Lexical HRNN in Table 2. Given the limited amount of TDT4 data, we found that the simpler DNN resulted in a higher F1 score on the test set. The fusion HRNNs in Table 2 are the results of the three fusion models, corresponding to Figure 2 (a,b,c). We observed improvements in F1 score for all fusion methods using acoustic features, especially when using statistical features and sentence embedding. We also applied fusion of the same acoustic statistics to the DNN model by concatenating the embedding vector with its input BOW features. However, as shown in the bottom row of Table 2, the DNN-based fusion resulted in a small reduction in F1 score.

3.3. Additional Text Training Data

We further evaluated on the same test data using a model trained on a joint set of TDT4 data and TDT2 data [38]. The TDT2 data was divided into 1469 training and 195 validation programs (see Table 1). When training the models, the corresponding acoustic features were set to zero for the TDT2 data. The story segments were shuffled within each TDT4 or TDT2 dataset as in Section 2.4, so that any adjacent stories both have either acoustic features or zero values for the parameter update.

The results are shown in Table 3. Using the additional TDT2 training data, the purely lexical HRNN resulted in an improved F1 score similar to that obtained with the fusion systems trained on TDT4. Further improvements were observed for the TDT4+TDT2 Fusion HRNNs, except the one using statistical features. In this experiment, the fusion model with the sentence embedding of acoustic features had the best F1 score, despite most of the acoustic features for training being set to 0.

3.4. Testing with ASR transcription

The experiments so far have used human transcriptions of the TDT4 data. In our final experiment we tested on the TDT4 test data using ASR transcriptions and automatic punctuation, rather than human transcription. We investigated training using both human and ASR transcriptions of the TDT 4 data, together with the text-only TDT2 data (human transcription).

We used the ASR system we developed for the transcription of British Broadcasting Corporation (BBC) TV data [39], using around 600 hours of training data taken from the 2015 Multi-Genre Broadcast (MGB) Challenge [40][2]. The system

[1]The F1-measure was computed with a tolerance window of 50 words according to the TDT2 standard [38].

[2]http://www.mgb-challenge.org

Table 4. F1-measure comparisons testing on TDT4 data using ASR transcriptions, with models trained with TDT4 data, using either human transcriptions or ASR, and additional TDT2 text data.

Model	F1-Measure	
	Human trans	ASR trans
Lexical DNN [18]	0.680	0.691
Lexical HRNN [31]	**0.689**	**0.704**
Fusion HRNN (statistics)	**0.706**	0.716
Fusion HRNN (sentence embedding)	0.704	**0.736**
Fusion HRNN (sentence gap embedding)	0.683	0.697
Fusion DNN (statistics)	0.669	0.698

is based on the sequence-trained deep neural networks in a hybrid configuration, following [41]. On the 2015 MGB Challenge development dataset, this system resulted in a word error rate (WER) of 28%. We note that the TDT4 data is mainly American English in contrast to the largely British English MGB data.

We then applied automatic punctuation [21] to the ASR transcription described above. The punctuation system segmented the ASR transcription splitting at pauses with a duration of over 0.2 seconds, followed by a neural machine translation model trained to map a word sequence to punctuation marks (full stop, comma, exclamation mark, question mark, ellipsis).

The results are shown in Table 4, where our hierarchical RNN model also exceeded the performance of the state-of-the-art with ASR transcriptions. By fusing acoustic features with the lexical features, we observed further improvement with the fusion models of both statistics and sentence embedding. Especially, sentence embedding fusion model was consistently effective for all the experiments in this paper. In this experiment, we did not gain any improvement with sentence gap embedding; the sentence gap embedding approach is more strongly dependent on the accuracy of automatic punctuation and sentence segmentation. Finally we note that training on ASR transcripts results in slightly higher F1 scores than training on human transcripts, possibly due to a better match with the test set.

4. CONCLUSIONS

This paper proposes addresses story segmentation using lexical and acoustic feature fusion with a hierarchical RNN model. The topic structure is captured using a hierarchical model based on an RNN sentence modeling layer and a bidirectional LSTM topic modeling layer. The two modalities are fused in the sentence level topic modeling layer. We train our model using the relatively low-resource TDT4 data containing audio and text transcriptions. We show that augmenting this training data with additional text data from other news sources (TDT2) helps to improve the precision of the system.

We conducted experiments comparing the combinations of lexical and acoustic features and combinations of training data. Experimental results on TDT4 test data indicated that the hierarchical RNN topic modeling can take advantage of the fusion of acoustic and lexical modalities, especially when additional text training data is available. Our fusion model using sentence embedding results in a higher F1 score for story segmentation when compared with conventional state-of-the-art methods, using both human and ASR transcriptions.

5. ACKNOWLEDGEMENTS

This work was supported by the EU H2020 project SUMMA, under grant agreement 688139.

6. REFERENCES

[1] Manabu Okumura and Takeo Honda, "Word sense disambiguation and text segmentation based on lexical cohesion," in *Proc. of COLING*, 1994, pp. 755–761.

[2] Marti A. Hearst, "TextTiling: Segmenting text into multi-paragraph subtopic passages," *Computational Linguistics*, vol. 23, no. 1, pp. 33–64, 1997.

[3] J.P. Yamron, I. Carp, L. Gillick, S.Lowe, and P. van Mulvregt, "A hidden Markov model approach to text segmentation and event tracking," in *Proc. of ICASSP*, 1998, vol. 1, pp. 333–336.

[4] Freddy Choi, "Advances in domain independent linear text segmentation," in *Proc. of the First Meeting of the North American Chapter of the Association for Computational Linguistics*, 2000, pp. 26–33.

[5] Martin Franz, Bhuvana Ramabhadran, Todd Ward, and Michael Picheny, "Automated transcription and topic segmentation of large spoken archives," in *Eurospeech*, 2003, pp. 953–956.

[6] Nicola Stokes, Joe Carthy, and Alan Smeaton, "Select: A lexical cohesion based news story segmentation system," *Journal of AI Communications*, vol. 17, no. 1, pp. 3–12, 2004.

[7] Minmi Lu, Lilei Zheng, Cheung-Chi Leung, Lei Xie, Bin Ma, and Haizhou Li, "Broadcast news story segmentation using probabilistic latent semantic analysis and Laplacian eigenmaps," in *Porc. of APSIPA ASC*, 2011, pp. 356–360.

[8] Elizabeth Shriberg, Andreas Stolcke, Dilek Hakkani-Tür, and Gökkan Tür, "Prosody-based automatic segmentation of speech in to sentences and topics," in *Speech Communication*, 2000, vol. 32, pp. 127–154.

[9] Pei-Yun Hsueh, Johanna D Moore, and Steve Renals, "Automatic segmentation of multiparty dialogue," in *EACL*, 2006.

[10] Andrew Rosenberg and Julia Hirschberg, "Story segmentation of broadcast news in English, Mandarin and Arabic," in *Proc. of HLT-NAACL*, 2006, pp. 125–128.

[11] Alexander Hauptmann and Michael Witbrock, "Story segmentation and detection of commercials in broadcast news video," in *Proc. of Advances in Digital Libraries*, 1999, pp. 168–179.

[12] Winston Hsu, Shih-Fu Chang, Chih-Wei Huang, Lyndon Kennedy, Ching-Yung Lin, and Giridharan Iyengar, "Discovery and fusion of salient multi-modal features towards news story segmentation," in *IS&-T/SPIE Electronic Imaging*, 2004.

[13] Winston Hsu, Lyndon Kennedy, Shih-Fu Chang, Martin Franz, and John R. Smith, "Columbia-IBM news video story segmentation in TRECVID 2004," Tech. Rep. 207-2005-3, Columbia University, 2005.

[14] 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.

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

[16] David M. Blei, Andrew Y. Ng, and Michael I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[17] Hemant Misra, Frank Hopfgartner, Anuj Goyal, P. Punitha, and Joemon M. Jose, "TV news story segmentation based on semantic coherence and content similarity," in *International Conference on Advances in Multimedia Modeling*, 2010, pp. 347–357.

[18] Jia Yu, Xiong Xiao, Lei Xie, Eng Siong Chng, and Haizhou Li, "A DNN-HMM approach to story segmentation," in *Proc. of Interspeech*, 2016, pp. 1527–1531.

[19] Yiping Song, Lili Mou, Rui Yan, Li Yi, Zinan Zhu, Xiaohua Hu, and Ming Zhang, "Dialogue session segmentation by embedding-enhanced TextTiling," in *Proc. of Interspeech*, 2016, pp. 2706–2710.

[20] Chenglin Xu, Lei Xie, Guangpu Huang, Xiong Xiao, Eng Siong Chng, and Haizhou Li, "A deep neural network approach for sentence boundary detection in broadcast news," in *Proc. of Interspeech*, 2014, pp. 2887–2891.

[21] Ondřej Klejch, Peter Bell, and Steve Renals, "Punctuated transcription of multi-genre broadcasts using acoustic and lexical approaches," in *Proc. of Spoken Language Technology Workshop*, 2016.

[22] Yoshua Bengio, Réjean Duchame, Pascal Vincent, and Christian Janvin, "A neural probabilistic language model," *Journal of Machine Learning Research*, , no. 3, pp. 1137–1155, 2003.

[23] Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur, "Recurrent neural network based language model," in *Proc. of Interspeech*, 2010, pp. 1045–1048.

[24] Martin Sundermeyer, Ralf Schlüter, and Hermann Ney, "LSTM neural network for language modeling," in *Proc. of Interspeech*, 2012, pp. 194–197.

[25] Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazzeer, and Yonghui Wu, "Exploring the limits of language modeling," *ArXiv:1602.02410*, 2016.

[26] Quoc Le and Tomas Mikolov, "Distributed representation of sentences and documents," in *Proc. of ICML*, 2014, pp. 1188–1196.

[27] Andrew Dai, Christopher Olah, and Quoc Le, "Document embedding with paragraph vector," in *Proc. of NIPS 2014 in Deep Learning and Representation Learning Workshop*, 2014.

[28] Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei, "Sharing clusters among related groups: Hierarchical Dirichlet process," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1566–1581, 2006.

[29] Justin Grimmer, "A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases," *Political Analysis*, vol. 18, no. 1, pp. 1–35, 2010.

[30] Rui Lin, Shujie Liu, Muyun Yang, Mu Li, Ming Zhou, and Sheng Li, "Hierarchical recurrent neural network for document modeling," in *Proc. of Empirical Methods in Natural Language Processing*, 2015, pp. 899–907.

[31] Emiru Tsunoo, Peter Bell, and Steve Renals, "Hierarchical recurrent neural network for story segmentation," in *Proc. of Interspeech*, 2017.

[32] Melissa Sherman and Yang Liu, "Using hidden Markov models for topic segmentation of meeting transcripts," in *Proc. of SLT*, 2008, pp. 185–188.

[33] Hamid Palangi, Li Deng, Yelong Shen, Jianfeng Gao, Xiaodong He, Jianshu Chen, Xinying Song, and Rabab Ward, "Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval," *CoRR, abs/1502.06922*, 2015.

[34] Alain de Cheveigne and Hideaki Kawahara, "YIN, a fundamental frequency estimator for speech and music," *The Journal of Acoustical Society of America*, pp. 1917–1930, 2002.

[35] George Karypis, "CLUTO - a clustering toolkit," Tech. Rep., Dept. of Computer Science, University of Minnesota, 2002.

[36] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio, "On the difficulty of training recurrent neural networks," *arXiv:1211.5063v2*, 2013.

[37] Stephanie Strassel, Jumbo Kong, and David Graff, "TDT4 multilingual text and annotations," 2003.

[38] Jon Fiscus, George Doddington, John Garofolo, and Alvin Martin, "NIST's 1998 topic detection and tracking evaluation (TDT2)," in *Proc. of DARPA Broadcast News Workshop*, 1999, pp. 19–24.

[39] Peter Bell, Catherine Lai, Clare Llewellyn, Alexandra Birch, and Mark Sinclair, "A system for automatic broadcast news summarisation, geolocation and translation," in *Proc. of Interspeech*, 2015.

[40] P Bell, MJF Gales, T Hain, J Kilgour, P Lanchantin, X Liu, A McParland, S Renals, O Saz, M Wester, and PC Woodland, "The MGB Challenge: Evaluating multi-genre broadcast media recognition," in *Proc IEEE ASRU*, 2015.

[41] Karel Veselý, Arnab Ghoshal, Lukás Burget, and Daniel Povey, "Sequence-discriminative training of deep neural network," in *Proc. of Interspeech*, 2013, pp. 2345–2349.