A HIERARCHICAL NEURAL SUMMARIZATION FRAMEWORK FOR SPOKEN DOCUMENTS

Tzu-En Liu[1], *Shih-Hung Liu*[3], *Berlin Chen*[1,2]

[1] National Taiwan Normal University, Taipei, Taiwan
[2] Pervasive Artificial Intelligence Research (PAIR) Labs, Taiwan
[3] Delta Management System, Taipei, Taiwan

ABSTRACT

Extractive text or speech summarization seeks to select indicative sentences from a source document and assemble them together to form a succinct summary, so as to help people to browse and understand the main theme of the document efficiently. A more recent trend is towards developing supervised deep learning based methods for extractive summarization. This paper extends and contextualizes this line of research for spoken document summarization, while its contributions are at least three-fold. First, we propose a neural summarization framework with the flexibility to incorporate extra acoustic/prosodic and lexical features, for which the ROUGE evaluation metric is embedded into the training objective function and can be optimized with reinforcement learning. Second, disparate ways to integrate acoustic features into this framework are investigated. Third, the utility of our proposed summarization methods and several widely-used state-of-the-art ones are extensively compared and evaluated. A series of empirical experiments seem to demonstrate the effectiveness of our summarization methods.

Index Terms— Extractive spoken document summarization, reinforcement learning, convolutional neural network, recurrent neural network, hierarchical encoding

1. INTRODUCTION

Speech is one of the most important sources of information about multimedia. By means of extractive speech summarization, which manage to select a concise set of salient sentences from a spoken document according to a target summarization ratio, people can efficiently browse and digest multimedia content by listening to the associated speech segments of a summary. Extractive speech summarization may inevitably suffer from the issue of incorrect information caused by recognition errors when using automatic speech recognition (ASR) techniques to transcribe spoken documents into text form. Nevertheless, extractive speech summarization also presents unique opportunities that do not exist for text summarization; for instance, inherent characteristics about prosody/acoustics and emotion/speakers may facilitate the determination of the important parts and structures of a spoken document. There is thus an urgent need to design and develop pragmatic extractive speech summarization methods for potential practical applications.

The wide range of extractive speech summarization methods developed to date can be broadly categorized into three groups: 1) methods simply based on sentence position or structure information, 2) methods based on unsupervised sentence ranking, and 3) methods based on supervised sentence classification. Apart from this, an extractive summary may also be generated by considering several other aspects like being generic or query-oriented summarization, single-document or multi-document summarization, and so forth. The interested reader is referred to [1] for a comprehensive overview of this area. We, in this paper, will concentrate exclusively on generic, extractive speech summarization because it usually constitutes the essential underpinning for many other speech summarization tasks.

On a separate front, recent years have seen a surge of interest in developing and applying supervised or unsupervised deep neural network-based methods to extractive text summarization. A general school of thought is to frame extraction summarization as a sequence label problem [2-8], where each sentence of a document to be summarized is quantified with a score (or tagged with a label) that help determine whether the sentence should be included in the final summary or not. Most of the current instantiations typically follow a two-step strategy: 1) a recurrent neural network (RNN)-based encoder is first employed to obtain a holistic representation of the document by taking the representations of its constituent sentences as the inputs to RNN successively; 2) an RNN-based decoder that takes the document representation as the initial input is then used to quantify (or label) each sentence of the document in tandem, meanwhile allowing for the previously processed sentences to avoid redundancy. This paper extends and contextualizes this line of research for spoken document summarization, while its contributions are at least three-fold. First, a neural summarization framework with the flexibility to incorporate extra acoustic/prosodic and lexical features is proposed, for which the ROUGE evaluation metric (*cf.* Section 4.1) is embedded into the training objective function and can be optimized with reinforcement learning. Second, we investigate disparate ways to inject the information of acoustic features into this framework. Third, the performance of our proposed summarization methods and several widely-used state-of-the-art ones are thoroughly compared and evaluated.

2. SUMMARIZATION MODELING FRAMEWORK

In this section, we introduce our proposed neural summarization modeling framework which consists of three main components: hierarchical encoder, decoder and reinforcement learning. Given a spoken document $D=\{S_1, S_2, \ldots, S_N\}$ (a sequence of spoken sentences), our model will select a subset of M sentences from D and in turn concatenate them to form the summary. Technically, for each spoken sentence $S_i \in D$, the model will predict a score for S_i in terms of $P(1|S_i,D,\theta)$, where $y_i \in \{0,1\}$ and 1 indicates that S_i

should be included in the summary (and 0 otherwise). We can use $P(1|S_i,D,\theta)$ as an indicator to judge whether S_i should be selected into the summary or not: the higher the value of $P(1|S_i,D,\theta)$, the more likely that S_i will be a summary sentence. After predicting the score of each spoken sentence, we then rank those sentences based on the scores $P(1|S_i,D,\theta)$ and pick up top-M sentences to produce the final summary. In what follows, we will elaborate on the architecture of the proposed neural summarization modeling framework and its main components (*cf.* Figure 1).

2.1. Hierarchical encoder

The hierarchical encoder is designed to have two encoding levels: one is the lower-level sentence encoder and the other is higher-level document encoder.

2.1.1. Sentence encoder

A convolutional neural network (CNN) can deal with text sentences of variable lengths and projects them into a low-dimensional vector space to obtain fixed-length vectors (or continuous representations). Many previous studies have also confirmed its utility for this purpose in an array of natural language processing (NLP) applications [2, 8-14]. In this paper, we thus adopt CNN to produce a fixed-length representation for every sentence of the spoken documents to be summarized. To do this, one-dimensional temporal convolution with a kernel filter K of width N is applied successively on a window of N words in a spoken sentence S_i to form a new feature map. Such a notion of sequentially capturing a text-span of N words bears some resemblance to the traditional N-gram language model for capturing the co-occurrence and proximity information of consecutive words in a short-span manner. Once the feature maps have been produced, max-pooling is in turn applied on those feature maps over time, taking the maximum values of their respective elements, which collectively form a fixed-length feature vector (representation) for the spoken sentence S. Multiple kernel filters of various sizes, where each kernel filter are applied multiple times with different parameter settings, can assist in constructing sentence representations for better task performance.

2.1.2. Document encoder

The document encoder compiles a sequence of sentences of a document to be summarized to produce a fixed-length document representation in a holistic manner. To this end, we employ a recurrent neural network (RNN) with long short-term memory (LSTM) cells, where LSTM can help alleviate the vanishing gradient problem when RNN is trained with long sequences of inputs [15]. In addition, following the common practice of implementation, we reverse the order of sentences to be fed into the document encoder [6,7,16]. By doing so, RNN can encapsulate more information about the sentences that are positioned at the beginning of the document to be summarized, since it is anticipated that the leading sentences would often contain more salient semantic content of the document. Formally, we use the following equations to obtain the document representation:

$$\mathbf{h}_i = f(\mathbf{h}_{i+1}, \mathbf{s}_i) \tag{1}$$

$$\mathbf{d} = \mathbf{h}_1 \tag{2}$$

where $f(.)$, $\mathbf{s}_i$, $\mathbf{h}_1$ and $\mathbf{d}$ stand for the LSTM function, sentence representation, representation of hidden layer output at first time stamp and document representation, respectively. It should be noted here that in order to obtain the representation of the hidden layer output $\mathbf{h}_i$, we need to feed in the representations of the sentences into RNN in reverse order.

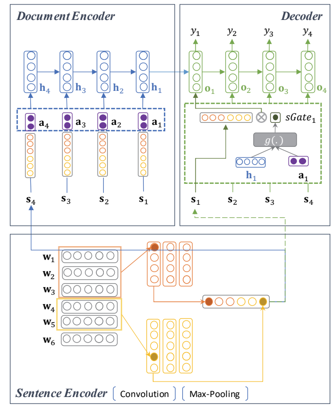

Figure 1. The proposed neural summarization framework, which consists of three components: 1) sentence encoder, 2) document encoder and 3) decoder. The acoustic features can be exploited by using concatenation in the document decoder (dashed-box in the upper-left) or by a weighting gate conditioned on the hidden layer representation h_i and the acoustic information $\mathbf{a}_i$, as shown in the upper-right dashed-box.

2.2. Decoder

The decoder is equipped with another RNN (with LSTM) and a softmax function to sequentially quantify each sentence S_i in the spoken document to be summarized D with a score (in terms of $P(1|S_i,D,\theta)$) ranging from 0 to 1 (1 signifies highly relevant, while 0 highly irrelevant to the document). The inputs of the decoder are the spoken sentence representations which are constructed previously from the sentence encoder. The order of the input sentences is along the normal direction which is different from that of the document encoder. We can use the following equation to obtain the representation of a hidden layer output for the decoder:

$$\mathbf{o}_i = f(\mathbf{o}_{i-1}, \mathbf{s}_i) \tag{3}$$

$$\mathbf{o}_0 = \mathbf{d} \tag{4}$$

where $\mathbf{o}_0$ and $\mathbf{o}_i$ denote the hidden layer representations at the initial time stamp and time stamp i, respectively. This way, the decoder is able to identify important sentences within the spoken document from both the local and global perspectives at the same time. Finally, we rank all spoken sentences based on the scores in terms of $P(1|S_i,D,\theta)$ which are produced by the softmax functions at the output layers of the decoder at different time stamps.

2.3. Reinforcement learning

A conventional supervised, extractive neural summarization model usually estimated by maximizing the product of the probabilities of the corresponding ground-truth labels of sentences involved in a training document that are predicted by the summarization model. This is equivalent to minimizing the cross-entropy loss at each time stamp of decoding, which will serve as the training objective

$$L(\theta) = -\sum_{i=1}^{n} \log P(y_i \mid S_i, D, \theta) \tag{5}$$

The cross-entropy loss would lead to significant incongruity between the training and test of a summarization model, since the ultimate evaluation metric (e.g., the variants of ROUGE measure; see Section 4.1) is not taken into account in the training objective function. Phrased another way, the summarization model trained based on Eq. (5) will aim to rank sentences with the maximum likelihood criterion of generating their summary labels, whereas the summary generated by the model is evaluated with a different metric (e.g., the variants of ROUGE measure) at the test phase.

In view of this, we reformulate the training of our neural summarization model with the reinforcement learning paradigm so as to alleviate such discrepancy. Reinforcement learning [17] introduces the so-called reward function $r(.)$ into the training objective function of the summarization model, which would make the objective function more closely coupled with the ultimate evaluation metric of extractive speech summarization. For the idea to work, the reward function $r(.)$ is embodied with the average score of the variants of ROUGE measure (*cf.* Section 4.1). Mathematically, the training objective function of reinforcement learning is minimizing the negative expected reward expressed by

$$L(\theta) = -\mathrm{E}_{\hat{\mathbf{y}} \sim p_\theta}[r(\hat{\mathbf{y}})] \tag{6}$$

where P_θ denotes $p(\cdot \mid D, \theta)$, the distribution of the summary labels of all sentences involved in a given training document D with parameter θ, and $\hat{\mathbf{y}} = \hat{y}_1, \ldots, \hat{y}_N$ is the predicted summary label sequence of the document D based on a certain sampling method. In estimation, since the reward function is non-differentiable, we thus rewrite the gradient of training objective function as follows:

$$\nabla L(\theta) \approx -r(\hat{\mathbf{y}}) \sum_{i=1}^{n} \nabla \log p(\hat{y}_i \mid S_i, D, \theta) \tag{7}$$

3. LEVERAGING ACOUSTIC FEATURES

In this section, we explain how to integrate acoustic features into the proposed neural summarization framework in three different ways, viz. integrating document-level acoustic features in the encoder, sentence-level acoustic features in the decoder and the indicator of sentence-level acoustic features in decoder. The dashed-boxes shown in Figure 1 illustrate the idea of integrating useful extra features in the proposed summarization framework. First, for the document-level acoustic features, we concatenate the sentence-as-unit acoustic feature vector $\mathbf{a}_i$ in the document encoder:

$$\mathbf{h}_i = f(\mathbf{h}_{i+1}, [\mathbf{s}_i; \mathbf{a}_i]) \tag{8}$$

The document representation will thus contain acoustic information encoded by RNN over time with Eq. (8). Second, for the sentence-level acoustic features used in the decoder, we also concatenate the sentence-as-unit acoustic information $\mathbf{a}_i$ with $\mathbf{s}_i$:

$$\mathbf{o}_i = f(\mathbf{o}_{i-1}, [\mathbf{s}_i; \mathbf{a}_i]) \tag{9}$$

The purpose of this consideration is that we can take the individual sentence-level acoustic features into account when performing the decoding process. Third, inspired from the notion of the so-called selective mechanism [18], we hope to exploit the acoustic features as an indicator to help select candidate summary sentences. Therefore, we design a weighting-gate based neural network which is conditioned on the hidden layer representation $\mathbf{h}_i$ and acoustic features $\mathbf{a}_i$ to modulate spoken sentences representations:

$$sGate_i = g(W_s[\mathbf{h}_i; \mathbf{a}_i] + \mathbf{b}) \tag{10}$$

Table 1. The statistical information of MATBN used in the summarization experiments.

	Training Set	Evaluation Set
Number of Doc.	185	20
Avg. Num. of Sent. per Doc.	20	23.3
Avg. Num. of words per Sent.	17.5	16.9
Avg. Num. of words per Doc.	326.0	290.3
Avg. Word Error Rate (WER)	38.0%	39.4%
Avg. Char. Error Rate (CER)	28.8%	29.8%

Table 2. Acoustic features of each spoken sentence.

Acoustic features	1. Pitch (min, max, diff, avg.) 2. Peak normalized cross-correlation of pitch (min, max, diff, avg.) 3. Energy value (min, max, diff, avg.) 4. Duration value (min, max, diff, avg.) 5. 1st formant value (min, max, diff, avg.) 6. 2nd formant value (min, max, diff, avg.) 7. 3rd formant value (min, max, diff, avg.)

where $g(.)$ is a three-layer feed-forward neural network. The value of $sGate_i$ is in the range from 0 to 1, and can be treated as the importance indicator of a spoken sentence. This way, we multiply the vector representation $\mathbf{s}_i$ of a spoken sentence S_i and $sGate_i$ to produce a new vector representation $\mathbf{s}'_i$ and then replace the original representation $\mathbf{s}_i$ to be in turn fed into the decoder:

$$\mathbf{s}'_i = sGate_i \cdot \mathbf{s}_i \tag{11}$$

$$\mathbf{o}_i = f(\mathbf{o}_{i-1}, \mathbf{s}'_i) \tag{12}$$

To recap, our neural summarization modeling framework is general and enables us to incorporate additional useful features (such as acoustic and/or linguistic features) into it for better summarization performance. Note here that if we remove the functionality of the dashed-boxes depicted in Figure 1, the resulting model boils down to the one previously descried in [6].

4. EXPERIMENTS

4.1. Experimental dataset

We conduct an extensive set of summarization experiments on a Mandarin Benchmark broadcast new corpus (MATBN) [19], which has been widely used to evaluate several natural language processing (NLP)-related tasks, including speech recognition [20], information retrieval [21] and summarization [2, 4, 8]. A subset of 205 broadcast news documents compiled between November 2001 and August 2002 was reserved for the summarization experiments. Furthermore, since broadcast news stories often follow a relatively regular structure as compared to other speech materials like conversations, the positional information would play an important role in extractive summarization of broadcast news stories. We hence chose 20 documents, for which the generation of reference summaries is less correlated with the positional information (or the position of sentences) as the held-out test set to evaluate the general performance of the models instantiated from our summarization framework, while the other subset of 185 documents in the training set alongside their respective human-annotated summaries for estimation of the various supervised summarization methods compared in the paper. Table 1 shows some basic statistics about the training and evaluation sets.

For the assessment of summarization performance, we adopt three common variants of the widely-used ROUGE measure (viz. ROUGE-1, ROUGE-2 and ROUGE-L) [22]. All the experimental results reported hereafter are obtained by calculating the F-scores of these variants. The summarization ratio, defined as the ratio of the number of words in the automatic (or manual) summary to that in the reference transcript of a spoken document, was set to 10% in this research. Table 2 shows the acoustic features used in this study, where the total number of features is 35.

4.2. Experimental results

At the beginning, we first assess the performance levels of several existing widely-practiced conventional unsupervised methods, viz. vector space model (VSM) and latent semantic analysis (LSA) [23], and well-practiced neural network-based unsupervised methods, viz. skip-gram (SG) [2, 24] and continuous bag-of-word (CBOW) [2, 24], as well as a few strong supervised neural summarizers, including three-layer deep neural network (DNN) [8], convolutional neural network (CNN) [8] and Refresh [6] for extractive spoken document summarization (with the best configurations respectively found in the above studies). The corresponding results of these methods are illustrated in Table 3, where TD denotes the results obtained based on the manual transcripts of spoken documents and SD denotes the results using the recognition transcripts that may contain speech recognition errors. Several conspicuous observations can be made from Table 3. First, when conducted in an unsupervised manner, the neural network-based methods (viz. SG and CBOW) always outperform the classical vector-based methods (viz. VSM and LSA) for both the TD and SD cases. Second, the supervised summarizers, including DNN, CNN and Refresh, perform better than SG and CBOW in the TD case; however, the situation seems reversed for the SD case. This probably indicates that imperfect speech recognition affects supervised summarizers (viz. DNN, CNN and Refresh) much more than the unsupervised ones (viz. SG and CBOW). Third, our best neural summarization model (denoted by NSM-LE; *cf.* Table 3) is superior to the competitive supervised neural network-based summarizers (DNN, CNN and Refresh) for both the TD and SD cases. In particular, the most basic form of our summarization models (denoted by NSM w/o acoustic features) outperform the Refresh method for both the TD and SD cases.

In the last set of experiments, we carry out several experiments to assess the performance gains of injecting the acoustic information into our NSM model in disparate ways (viz. NSM-GE, NSM-LE, NSM-GE+LE and NSM-SE; these acronyms are explained in the bottom part of Table 3). Their corresponding results are highlighted with the gray background in Table 3. As can be seen, the best result is achieved by NSM-LE. A possible explanation is that since the acoustic features can be directly referenced during decoding, so that more complete acoustic information cues can be employed to assist in decoding better summary sentences. Furthermore, in the situation of using GE (including NSM-GE and NSM-GE+LE), our NSM model is less effective that those without using acoustic features (i.e. NSM w/o acoustic features) for the TD case, but it instead performs better for the SD case. This is probably because that the SD case suffers from speech recognition errors, where the acoustic features can to some extent compensate for such an undesirable effect. As a final note, we have also explored using the so-called gated recurrent unit (GRU) [25] in replace of LSTM for the RNN modeling in our hierarchical neural summarization framework, which shows competitive summarization results with less training time in comparison to the latter. Here we, however, omit the detailed results of using GRU due to space limitations.

Table 3. Summarization results achieved by our proposed method and several well-practiced methods.

		ROUGE-1	ROUGE-2	ROUGE-L
TD	VSM	0.347	0.228	0.290
	LSA	0.362	0.233	0.316
	SG	0.410	0.300	0.364
	CBOW	0.415	0.308	0.366
	DNN	0.488	0.382	0.444
	CNN	0.501	0.407	0.460
	Refresh	0.453	0.372	0.446
	NSM w/o acoustic features	0.470	0.388	0.459
	NSM-GE	0.458	0.383	0.448
	NSM-LE	**0.524**	**0.453**	**0.516**
	NSM-GE+LE	0.440	0.341	0.428
	NSM-SE	0.499	0.415	0.487
SD	VSM	0.342	0.189	0.287
	LSA	0.345	0.201	0.301
	SG	0.378	0.239	0.333
	CBOW	0.393	0.250	0.349
	DNN	0.371	0.233	0.332
	CNN	0.370	0.208	0.312
	Refresh	0.329	0.197	0.319
	NSM w/o acoustic features	0.399	0.313	0.388
	NSM-GE	0.416	**0.336**	0.407
	NSM-LE	**0.420**	0.332	**0.410**
	NSM-GE+LE	0.407	0.321	0.395
	NSM-SE	0.378	0.286	0.366

Refresh：strong baseline model used in [6].
w/o acoustic features：our model without using acoustic features.
GE： document representations augmented with acoustic features.
LE： sentence representations augmented with acoustic features.
GE+LE：both sentence and document representations augmented with acoustic features.
SE： sentence representations weighted by acoustic features' indication.

5. CONCLUTIONS AND FUTURE WORK

In this paper, we have proposed an effective extraction-based neural summarization framework for spoken document summarization. The summarization evaluation metric (viz. the average of the variants of the ROUGE measure) is plugged into the training objective function which can be optimized with reinforcement learning. In addition, we explore three different ways to seamlessly integrate acoustic features into our proposed neural summarization model. The experimental results demonstrate that our methods indeed can lead to significant improvements.

6. ACKNOWLEDGEMENT

This research is supported in part by Chunghwa Telecom Laboratories under Grant Number TL-107-8201, and by the National Science Council, Taiwan, under Grant Number MOST 107-2634-F-008-004- through Pervasive Artificial Intelligence Research (PAIR) Labs, Taiwan.

7. REFERENCES

[1] J.-M. Torres-Moreno (Eds.), Automatic Text Summarization, WILEY-ISTE, 2014.

[2] K.-Y Chen *et al.*, "An Information Distillation Framework for Extractive Summarization," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 1, pp. 161–170, January 2018.

[3] J. Cheng and M. Lapata, "Neural Summarization by Extracting Sentences and Words," in Proc. ACL, pp. 484–494, 2016.

[4] S.-H. Liu *et al.*, "Combining Relevance Language Modeling and Clarity Measure for Extractive Speech Summarization," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 6, pp. 957–969, 2015.

[5] R. Nallapati *et al.*, "SummaRnNNer: A Recurrent Neural Network Based Sequence Model for Extractive Summarization of Documents," in Proc. AAAI, pp. 3075–3081, 2017.

[6] S. Narayan *et al.*, "Ranking Sentences for Extractive Summarization with Reinforcement Learning," in Proc. NAACL, pp. 1747–1759, 2018.

[7] S. Narayan *et al.*, "Neural Extractive Summarization with Side Information," CoRR abs/1704.04530, 2017.

[8] C.-I. Tsai *et al.*, "Extractive Speech Summarization Leveraging Convolutional Neural Network Techniques," in Proc. SLT, pp.13–16, 2016.

[9] R. Collobort *et al.*, "Natural Language Processing (almost) from Scratch," Journal of Machine Learning Research, vol. 12, pp. 2493–2537, 2011.

[10] N. Kalchbrenner *et al.*, "A Convolutional Neural Network for Modeling Sentences," in Proc. ACL, pp. 655–665, 2014.

[11] Y. Kim, "Convolutional Neural Networks for Sentence Classification," in Proc. EMNLP, pp. 1746–1751, 2014.

[12] Y. Kim *et al.*, "Character-aware Neural Language Models," in Proc. AAAI, pp. 2741–2749, 2016.

[13] T. Lei *et al.*, "Molding CNNs for Text: Non-linear, Non-consecutive Convolutions," in Proc. EMNLP, pp. 1565–1575, 2015.

[14] X. Zhang *et al.*, "Character-level Convolutional Networks for Text Classification," Advances in Neural Information Processing Systems 28, pp. 649–657, 2015.

[15] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[16] I. Sutskever *et al.*, "Sequence to Sequence Learning with Neural Networks," Advances in Neural Information Processing Systems 27, pp. 3104–3112, 2014.

[17] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, MIT Press, Cambridge, MA, 2017.

[18] Q. Zhou *et al.*, "Selective Encoding for Abstractive Sentence Summarization," in Proc. ACL, pp. 1095–1104, 2017.

[19] H.-M. Wang *et al.*, "MATBN: A Mandarin Chinese Broadcast News Corpus," International Journal of Computational Linguistics & Chinese Language Processing, vol. 10, no. 2, pp. 219–236, 2005.

[20] J.-T. Chien, "Hierarchical Pitman-Yor-Dirchlet Language Model," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 8, pp. 1259–1272, 2015.

[21] C. L. Huang and C. H. Wu, "Spoken Document Retrieval Using Multilevel Knowledge and Semantic Verification," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 15, no. 8, pp. 2551–2590, 2007.

[22] C. Y. Lin, "ROUGE: Recall-oriented Understudy for Gisting Evaluation," 2003 [Online]. Available: http://haydn.isi.edu/ROUGE/.

[23] Y. Gong and X. Liu, "Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis," in Proc. SIGIR, pp. 19–25, 2001.

[24] T. Mikolov *et al.*, "Distributed Representations of Words and Phrases and Their Compositionality," in Proc. ICLR, pp. 1–9, 2013.

[25] K. Cho et al., "Learning Phrase Representations Using RNN Encoder–Decoder for Statistical Machine Translation," in Proc. EMNLP, pp. 1724–1734, 2014.