SEMANTIC INFERENCE BASED ON NEURAL PROBABILISTIC LANGUAGE MODELING FOR SPEECH INDEXING

Chien-Lin Huang, Chiori Hori, Hideki Kashioka

National Institute of Information and Communications Technology, Kyoto, Japan
{chien-lin.huang, chiori.hori, hideki.kashioka}@nict.go.jp

ABSTRACT

This study presents a novel approach to spoken document retrieval based on neural probabilistic language modeling for semantic inference. The neural network based language model is applied to estimate word association in a continuous space. The different kinds of weighting schemes are investigated to represent recognized words of a spoken document into an indexing vector. The indexing vector is transferred into the semantic indexing vector through the neural probabilistic language model. Such a semantic word inference and re-weighting make the semantic indexing vector a suitable representation for speech indexing. Experimental results conducted on Mandarin Chinese broadcast news show that the proposed approach can achieve a substantial and consistent improvement of spoken document retrieval.

Index Terms— Speech indexing, semantic inference, spoken document retrieval, neural network, language model

1. INTRODUCTION

Multimedia indexing and retrieval are active fields of research in the information era [1]. The media content is including text, image, video, and speech. Speech is the most convenient way for the interaction of human-to-human and human-to-machine. Applications of spoken document retrieval (SDR) in entertainment, education, and business are rapidly growing. Due to the nature of speech, the spoken query and documents are difficult to directly compare in signals. The automatic speech recognition (ASR) system is conventionally applied to transcribe the spoken document into text for speech indexing and retrieval [2]–[7]. The main challenge of SDR is the imperfect and unstructured dictation result. The overlapping multigram phone sequences have been estimated for variable-length syllables indexing to solve problems of deletion and insertion errors on speech recognition [2]. Other subword-based speech indexing schemes were successfully proposed to address the out-of-vocabulary problem [3, 4]. The multi-level knowledge indexing was proposed to combine different information sources [5]. Approaches like, Query expansion, semantic inference, and relevance feedback are considered text retrieval methods to increase the informative and descriptiveness of data [8]–[10]. These text retrieval techniques provide good solutions for speech indexing.

Two types of retrieval models, vector space model [11] and probabilistic model [12], have been applied to measure the relevance between a query and a document based on occurrences of query terms. In the probabilistic model, language model (LM) methods to retrieval have been shown to perform well empirically [13, 14]. The conventional language model is derived from the discrete n-gram count. Although language model systems perform well under the n-gram count scheme, they are discrete distributions and suffer from unseen patterns. As a result, back-off and interpolation smoothing strategies are commonly used for unseen words. Recently, neural network (NN) based language models have been used in speech recognition [15] and show several benefits. In contrary to the conventional n-gram count language model, the smoothing of neural network based language models is applied in an implicit way. Due to the projection of the entire vocabulary into a small hidden layer, semantically similar words get clustered and thus this characteristic is very suitable for inference in speech indexing. Words get substituted by other words which are learned in the neural network to be related, while no such relation could be found in the conventional n-gram count using the original sparse training data [16]. Instead of counting words, the neural network based language model is in a continuous space to estimate probabilities [17].

In this paper, we describe a new speech indexing method based on neural probabilistic language modeling for semantic inference. Based on the vector space model, the neural probabilistic language model is used to estimate word association for semantic inference. For instance, the word "student" may relate to words "teacher" and "school" in semantic inference. Once the specific word is mis-recognized in the spoken document, we can find it with related words [18, 19]. In such a neural probabilistic language model, we explore semantic relations between words which are considered in neural networks to alleviate deletion and substitution problems of transcripts. While the present study is related to recent SDR approaches [2]–[7], it capitalizes on semantic inference using the neural network

978-1-4799-0356-6/13/$31.00 ©2013 IEEE

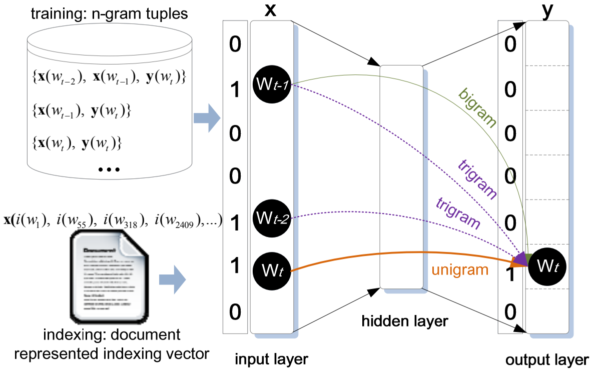
training: n-gram tuples
$\{\mathbf{x}(w_{t-2}),\ \mathbf{x}(w_{t-1}),\ \mathbf{y}(w_t)\}$
$\{\mathbf{x}(w_{t-1}),\ \mathbf{y}(w_t)\}$
$\{\mathbf{x}(w_t),\ \mathbf{y}(w_t)\}$
...
$\mathbf{x}(i(w_1),\ i(w_{55}),\ i(w_{318}),\ i(w_{2409}),\ldots)$
indexing: document represented indexing vector
x
y
0 1 0 0 1 1 0
Wt-1
Wt-2
Wt
bigram
trigram
trigram
unigram
0 0 0 0 0 1 0
hidden layer
input layer
output layer

Fig. 1. Architecture of the neural probabilistic language modeling for speech indexing.

based language model for speech indexing, which was not considered in earlier studies. The effectiveness of the proposed approach has been verified in experiments based on 1,550 anchor news stories collected from Mandarin Chinese broadcast news of 198 hours. In the following, we present the proposed speech indexing using the neural probabilistic language modeling for semantic inference in Section 2. Section 3 shows experiments in detail. We conclude with a summary of findings in Section 4.

2. THE NEURAL PROBABILISTIC LANGUAGE MODELING FOR SEMANTIC INFERENCE

2.1. Word Association for Semantic Inference

In general, word association can be established with concatenated or distant words in language models [17]. For example, the word bigram score $P(w_t \mid w_{t-1})$ is used to estimate the concatenation probability of a word sequence. Inspired by the neural network based language model [15, 20], this study proposed the semantic inference by finding word association and suggesting word expansion for speech indexing. With the word expansion and re-weighting, the semantic inference is expected to alleviate the problems of spoken document retrieval resulting from speech recognition errors. First, the neural probabilistic language model which reflects the global and continuous word association is made for semantic inference as shown in Fig. 1. In the training stage, n-gram words are coded into the control input vector $\mathbf{x}$ and the resulting output vector $\mathbf{y}$, respectively. The size of $\mathbf{x}$ and $\mathbf{y}$ is equal to the vocabulary size. When the neural network is learning with a bigram tuple $\{\mathbf{x}(w_{t-1}),\ \mathbf{y}(w_t)\}$, only values of neurons $\mathbf{x}(w_{t-1})$ and $\mathbf{y}(w_t)$ are set to 1 and others are 0. Both of $\mathbf{x}$ and $\mathbf{y}$ are sparse vectors. In the indexing stage, each document is represented as an indexing vector $\mathbf{x}$ by using the word weighting scheme. Then, the semantic indexing $\mathbf{y}$ can be estimated based on the indexing vector $\mathbf{x}$ and the learned neural network.

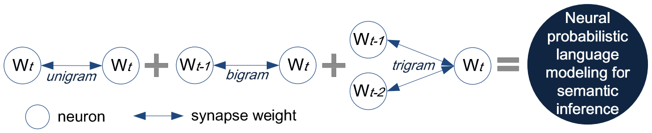
Wt
unigram
Wt
Wt-1
bigram
Wt
Wt-1
Wt-2
trigram
Wt
Neural probabilistic language modeling for semantic inference
neuron
synapse weight

Fig. 2. Example of n-gram word association in the neural probabilistic language modeling.

2.2. Activation and Learning

Fig. 2 demonstrates the concept of word association in the neural probabilistic language modeling. Neural networks can be viewed as weighted directed graphs in which neurons are nodes and directed edges (with weights) are connections between input and output neurons [21, 22]. The n-gram word association is modeled in the neural network as shown in Fig. 2. The sequential n-gram tuples of trigram $\{\mathbf{x}(w_{t-2}),\ \mathbf{x}(w_{t-1}),\ \mathbf{y}(w_t)\}$, bigram $\{\mathbf{x}(w_{t-1}),\ \mathbf{y}(w_t)\}$, and unigram $\{\mathbf{x}(w_t),\ \mathbf{y}(w_t)\}$ are applied to learn the NN based LM. The purpose of sequential n-gram tuples is to embed word association of unigram, bigram, trigram information into the neural network, and to estimate the global word association with n-gram connection patterns.

Based on these connection patterns, neural networks can be grouped into two categories including feed-forward and recurrent networks [22]–[25]. The loops occur because of feedback connections in recurrent networks but no loops are in feed-forward networks. In this study, we obtain the neural network language model by using the back-propagation learning based on feed-forward architectures with input, hidden and output layers. The values of neurons are estimated as follows:

$$y_j = f(\sum_i x_i \times a_{ji} + b_j) \tag{1}$$

where a_{ji} denotes the synapse weight from node i to node j in the neural network. b_j is the bias. The activation function is given by $f(.)$. The continuous sigmoid function is by far the most frequently used in neural networks because the cumulative distribution functions for many common probability distributions are like sigmoidal. It is a strictly increasing function which exhibits smoothness and has the desired asymptotic properties [22]. The standard sigmoid function is the logistic function defined by:

$$f(net_j) = \frac{1}{1+\exp^{-\beta\times net_j}} \tag{2}$$

where β is a slope parameter and $net = \mathbf{x}\cdot\mathbf{A}+\mathbf{b}$. $\mathbf{A}$ is the weight matrix, and $\mathbf{b}$ means the set of biases. Let $\{(\mathbf{x}^1,\mathbf{y}^1),(\mathbf{x}^2,\mathbf{y}^2),\ldots,(\mathbf{x}^o,\mathbf{y}^o)\}$ be a set of o n-gram training tuples, the back-propagation is a gradient decent method to minimize the squared-error cost function estimated as:

$$E = \frac{1}{2}\sum_{i=1}^{o}\left\|\mathbf{y}^i - \mathbf{d}^i\right\|^2 \quad (3)$$

where $\mathbf{y}$ and $\mathbf{d}$ indicate the desired and estimated output vectors, respectively. The key point in the proposed NN based LM is how the parameter $\lambda = \{a_{ji}, b_j\}$ over word association is estimated for semantic indexing. After learning is complete, the semantic inference depends on the NN based LM and spoken document representation for indexing, so we explore types of weighting schemes.

2.3. Document Representation and Semantic Indexing

In the a-bag-of-words framework, recognized words in a spoken document are computed as the indexing vector $\mathbf{x}$ by using the word weighting scheme $i(w_{k,d})$ in the indexing stage. $w_{k,d}$ denotes the word w in the document *d*. *k* means the dimension of the indexing vector. For semantic inference, the semantic indexing vector $\mathbf{y}$ is estimated based on neural networks with the learned parameter λ . The concept of the weighting scheme is the words which have low frequency in a document and occur in few documents will be discarded. We investigate spoken document representation by using three kinds of word weighting schemes.

TF-IDF: One popular weighting scheme is term frequency and inverse document frequency (TF-IDF),

$$i_{TF-IDF}(w_{k,d}) = u(w_{k,d}) \times v(w_{k,d}) \quad (4)$$

where $u(w_{k,d})$ is the term frequency weighting defined by:

$$u(w_{k,d}) = \frac{tf(w_{k,d})+1}{n_d} \quad (5)$$

where $tf(w_{k,d})$ represents the number of occurrences of the word $w_{k,d}$. n_d is a normalization which shows the total number of words in the document *d*. In addition, $v(w_{k,d})$ is the inverse document frequency computed as:

$$v(w_{k,d}) = \log(\frac{D}{df(w_{k,d})+1}) \quad (6)$$

where $df(w_{k,d})$ is the number of documents that contains at least one occurrence of the word $w_{k,d}$ in the database. *D* is the total number of documents.

Okapi BM25: The Okapi BM25 [26] is similar to TF-IDF but takes account of the document length for the term frequency computation in the word weighting scheme. The Okapi BM25 is defined by:

$$i_{BM25}(w_{k,d}) = \frac{tf(w_{k,d}) \times (\kappa+1)}{tf(w_{k,d}) + \kappa \times (1-c+c \times nl(d))} \times v(w_{k,d}) \quad (7)$$

where $nl(d)$ is the normalization document length. κ and c are empirically selected parameters.

Entropy: The third weighting scheme is the entropy estimation. Derived from TF-IDF, we replace IDF with the global entropy weighting as follows:

$$i_{entropy}(w_{k,d}) = u(w_{k,d}) \times [-\frac{df(w_{k,d})+1}{D} \cdot \log(\frac{df(w_{k,d})+1}{D})] \quad (8)$$

Our purpose is to estimation the significant of the specific word w_k in the document *d*. Since imperfect speech recognition results and the redundant transcription, not all of the recognized words are valid and meaningful. The word weighting scheme $i(w_{k,d})$ is used to eliminate those noises, emphasize semantic words, and has a better way of dealing with documents of different lengths. TF-IDF, Okapi BM25, and entropy weighting schemes are investigated for spoken document representation as the input indexing vector $\mathbf{x}$ and can further refer to the learned neural probabilistic language model as producing an output semantic indexing vector $\mathbf{y}$. The semantic indexing vector $\mathbf{y}$ denotes the semantic inference result for spoken document retrieval. We truncated the estimation by ignoring all words having a probability less than 0.0001 to improve the recall. Note that values of vectors $\mathbf{x}$ and $\mathbf{y}$ are either 0 or 1 in the training stage, which denote word association. But for the indexing stage, vectors of $\mathbf{x}$ and $\mathbf{y}$ can be any value between 0 and 1, which are the spoken document representation.

2.4. Spoken Document Retrieval

In spoken document retrieval, we treat the output semantic indexing vector $\mathbf{y}$ of neural networks as a vector space model (VSM) [11]. The similarity between the query $\mathbf{y}_q$ and spoken documents $\mathbf{y}_d$ is simply estimated by using a cosine measure:

$$\cos(\mathbf{y}_q, \mathbf{y}_d) = \frac{\mathbf{y}_q \times \mathbf{y}_d}{|\mathbf{y}_q| \cdot |\mathbf{y}_d|} = \frac{\sum_{k=1} \mathbf{y}_q(w_k) \times \mathbf{y}_d(w_k)}{\sqrt{\sum_{k=1} \mathbf{y}_q(w_k)^2} \times \sqrt{\sum_{k=1} \mathbf{y}_d(w_k)^2}} \quad (9)$$

In the cosine measure, $\cos(\mathbf{y}_q, \mathbf{y}_d)$, a value close to 1 means that two vectors are similar, whereas a value near 0 denotes

Table 1. F-score of top 5, 10, 15, and 20 retrieved documents for word indexing and proposed semantic indexing systems

Indexing/Top-n	5	10	15	20	25
Word(baseline)	44.45%	60.90%	61.77%	58.00%	53.24%
NN unigram	44.93%	61.10%	61.55%	57.81%	52.93%
NN bigram	46.46%	61.14%	62.17%	58.34%	53.87%
NN trigram	46.19%	58.76%	59.95%	56.49%	52.53%

two vectors are dissimilar. The retrieval results are ranked based on the similarity scores.

3. EXPERIMENTS

3.1. Experimental Framework

Experiments were reported based on Mandarin Chinese broadcast news MATBN database. MATBN was collected by Academia Sinica, Taiwan [27], which contained a total of 198 hours of broadcast news. 1,550 anchor news stories ranging over three years were extracted for experiments. The average document length of MATBN is 51.85 words. The word accuracy is 78.95%. We applied two standard evaluation metrics to evaluate the retrieval performance including F-score and mean average precision (mAP) [28]. 164 keyword queries (from two to four Chinese characters) were used in this study. The average length of queries is 3.02 Chinese characters. There are 15.71 relevant spoken documents in MATBN database.

3.2. Experimental Results

First, an objective F-score evaluation is employed to unigram, bigram, and trigram based neural networks. The speech indexing with the entropy weighting scheme was used and documents ranked at top 5, 10, 15, 20 and 25 as shown in Table 1. The baseline is the conventional word indexing system without applying semantic inference. The sequential n-gram tuples of trigram, bigram, and unigram are applied to learn NN based LMs. We denote the proposed semantic indexing systems are NN trigram, NN bigram and NN unigram in Table 1. The proposed semantic indexing systems perform well on the top 5 document ranking retrieval. Especially, NN bigram and NN trigram systems indicate the great improvement. The NN bigram system outperforms NN trigram, NN unigram, and baseline word indexing systems. The best F-score is achieved on the NN bigram system with the top 15 retrieved documents.

Second, Fig. 3 illustrates that mAP evaluations of TF-IDF, Okapi BM25, and Entropy weighting schemes compared with the conventional word indexing, NN unigram and NN bigram indexing systems. The NN bigram indexing system shows significant gains on different weighting schemes. Results of manual transcripts can be referred as the upper-bound of spoken document retrieval using ASR results. To know the effect of the proposed neural probabilistic language modeling on the text retrieval, the indexing by perfect manual transcripts was evaluated and noted as Entropy(M). The NN bigram system using semantic inference increases 0.15 mAP but no significant gains are found with clean text. Since the NN based LM preserves word association which reflects the expected semantic inference, the NN bigram indexing system offers around 2% absolute improvements over the baseline word index system, and achieves consistent improvements over evaluations of different word weighting schemes, mAP and F-score metrics, clean text and speech.

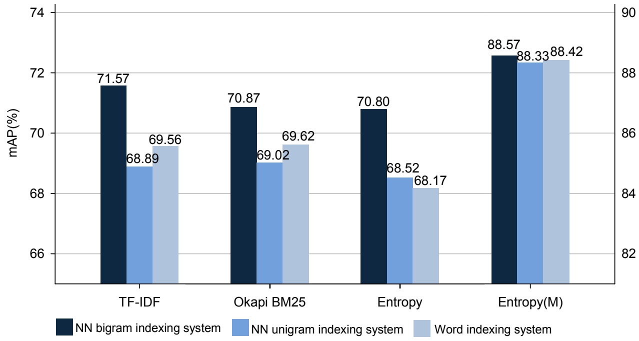

Fig. 3. mAP evaluations of various weighing schemes with conventional word indexing, NN unigram and bigram indexing.

4. CONCLUSION

In this article, we have presented a novel approach to spoken document retrieval using semantic inference. The speech recognition error is a critical issue in speech indexing. The neural probabilistic language modeling is used to explore word co-occurrence and association in a continuous space. Word association is applied for semantic indexing which is a way of inference to alleviate mis-recognition problems of ASR in SDR. We have investigated a variety of word weighting schemes on the semantic indexing and show comparisons of speech and text retrieval. Experimental results confirm that the proposed methods outperform the conventional word indexing. Although experiments were conducted on Mandarin Chinese broadcast news, the approach can be applied to other language and corpus. There are many directions for further research in exploring the use of deep belief networks and recurrent neural networks for language modeling. The relation between big data and n-gram selection is an interesting research issue. Implementing such neural network LMs is time-consuming with high computation cost. The efficient estimation appears to play an important role in applications.

5. ACKNOWLEDGMENTS

The authors would like to thank Bin Ma and Chung-Hsien Wu for their indispensable advice, support, and comments.

6. REFERENCES

[1] M. Lew, N. Sebe, C. Djeraba, and R. Jain, “Content-Based Multimedia Information Retrieval: State of the Art and Challenges,” *ACM Trans. Multimedia Computing, Communications, and Applications*, vol. 2, no. 1, pp. 1–19, 2006.

[2] K. Ng, “Subword-based Approaches for Spoken Document Retrieval,” Ph.D. dissertation, Mass. Inst. Technol., Cambridge, MA, 2000.

[3] P. Yu, K. Chen, C. Ma, and F. Seide, “Vocabulary-Independent Indexing of Spontaneous Speech,” *IEEE Trans. Speech Audio Processing*, vol. 13, no. 5, pp. 635–643. 2005.

[4] B. Logan, J.-M. Van Thong, and P. J. Moreno, “Approaches to Reduce the Effects of OOV Queries on Indexed Spoken Audio,” *IEEE Trans. Multimedia*, vol. 7, no. 5, pp. 899–906. 2005.

[5] C.-L. Huang and C.-H. Wu, “Spoken Document Retrieval Using Multi-Level Knowledge and Semantic Verification,” *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2551–2560, 2007.

[6] C. Chelba, T. J. Hazen, and M. Saraçlar, “Retrieval and Browsing of Spoken Content,” *IEEE Signal Processing Magazine*, vol. 24, no. 3, pp. 39–49, 2008.

[7] Y.-C. Pan, H.-Y. Lee, and L.-S. Lee, “Interactive Spoken Document Retrieval With Suggested Key Terms Ranked by a Markov Decision Process,” *IEEE Trans. Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 632–645, 2012.

[8] H. Cui, R. Sun, K. Li, M.-Y. Kan, and T.-S. Chua, “Question Answering Passage Retrieval Using Dependency Relations,” in *Proc. ACM SIGIR Conf.*, pp. 400–407, 2005.

[9] R. Wilkinson and P. Hingston, “Using the Cosine Measure in a Neural Network for Document Retrieval,” in *Proc. ACM SIGIR Conf.*, pp. 202–210, 1991.

[10] J. Rocchio, “Relevance Feedback in Information Retrieval,” in *The SMART Retrieval System: Experiments in Automatic Document Processing*, pp. 313–323. Prentice-Hall Inc., 1971.

[11] C. Buckley and J. Walz, “SMART in TREC 8,” in *Proc. Eighth Text REtrieval Conf. (TREC-8 "99), NIST Special Publication 500–264, Voorhees and Harman, eds.*, pp. 577–582, 2000.

[12] T. K. Chia, K. C. Sim, H. Li, and H. T. Ng, “A Lattice-Based Approach to Query-by-Example Spoken Document Retrieval,” in *Proc. ACM SIGIR Conf.*, pp. 363–370, 2008.

[13] C. Zhai and J. Lafferty, “A Study of Smoothing Methods for Language Models Applied to Ad Hoc Information Retrieval,” in *Proc. ACM SIGIR Conf.*, pp. 334–342, 2001.

[14] J. M. Ponte and W. B. Croft, “A Language Modeling Approach to Information Retrieval,” in *Proc. ACM SIGIR Conf.*, pp. 275–281, 1998.

[15] S. Kombrink, T. Mikolov, M. Karafi´at, and L. Burget, “Recurrent Neural Network based Language Modeling in Meeting Recognition,” in *Proc. Interspeech*, pp. 2877–2880, 2011.

[16] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, “A Neural Probabilistic Language Model,” in *Journal of Machine Learning Research*, vol. 3, pp. 1137–1155, 2003.

[17] G. Saon and J.-T. Chien, “Large-Vocabulary Continuous Speech Recognition Systems: A Look at Some Recent Advances,” *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 18–33, 2012.

[18] C.-L. Huang, B. Ma, H. Li, and C.-.H Wu, “Speech Indexing Using Semantic Context Inference,” in *Proc. Interspeech*, pp. 717–720, 2011.

[19] D. Song and P. D. Bruza, “Towards Context Sensitive Information Inference,” *J. Am. Soc. Information Science and Technology*, vol. 54, no. 4, pp. 321-334, 2003.

[20] T. Mikolov, S. Kombrink, L. Burget, J. ˇCernock´y, and S. Khudanpur, “Extensions of Recurrent Neural Network Language Model,” in *Proc. ICASSP*, pp. 5528–5531, 2011.

[21] S. Bengio and Y. Bengio, “Taking on the Curse of Dimensionality in Joint Distributions Using Neural Networks,” *IEEE Trans. Neural Networks*, vol. 11, no. 3, pp. 550–557. 2000.

[22] A. K. Jain and J. Mao, “Artificial Neural Networks: A Tutorial,” *IEEE Computer*, vol. 29, no. 3, pp. 31–44, 1996.

[23] J. L. Elman, “Finding Structure in Time,” *Cognitive Science*, vol. 14, pp. 179–211, 1990.

[24] J. B. Pollack, “The Induction of Dynamical Recognizers,” *Machine Learning*, vol. 7, pp. 227–252, 1991.

[25] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T.N. Sainath, and B. Kingsbury, “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.

[26] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford, “Okapi at TREC-3,” in *Proc. of the third Text REtrieval Conference (TREC-3)*, pp. 109–126, 1995.

[27] H.-M. Wang, B. Chen, J.-W. Kuo, and S.-S. Cheng, “MATBN: A Mandarin Chinese Broadcast News Corpus,” *Int. J. Comput. Ling. Chinese Lang. Process.*, vol. 10, no. 2, pp. 219–236, 2005.

[28] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*. New York: McGraw-Hill, 1983.