GENERATING PSEUDO-RELEVANT REPRESENTATIONS FOR SPOKEN DOCUMENT RETRIEVAL

Zheng-Yu Wu, Li-Phen Yen, Kuan-Yu Chen

National Taiwan University of Science and Technology, Taiwan

ABSTRACT

Spoken document retrieval (SDR) has become an important research subject due to the immenseness of multimedia data along with speech have spread around the world in our daily life. One of the fundamental challenge facing SDR is that the input query usually contains only a few words, which is too short to convey the information need of a user. In order to mitigate the problem, a well-practiced strategy is to reformulate the original query by performing a pseudo-relevance feedback process. Although several studies have evidenced its ability and capability for enhancing the retrieval performance, the time-consuming problem makes it hard to be used in reality. Motivated by the observations, in this paper, we concentrate on proposing a novel framework, which targets at generating a set of pseudo-relevant representations for a given query automatically, and eliminating the time-wasting problem. On top of the generated representations, we further investigate a novel query reformulation mechanism so as to improve the retrieval performance. A series of empirical SDR experiments conducted on a benchmark collection demonstrate the good efficacy of the proposed framework, compared to several existing strong baseline systems.

Index Terms— Spoken document retrieval, pseudo-relevance feedback, query reformulation, representation

1. INTRODUCTION

Owing to large volumes of multimedia data associated with spoken documents made available to the public, spoken content analysis has become an attractive and rising research subject over the past two decades in the speech processing community [1-3]. There are two main streams of research on processing a given text/spoken query and a spoken document. On one hand, spoken term detection (STD) [4, 5] embraces the goal of extracting probable spoken terms or phrases inherent in a spoken document that could match the query words or phrases literally; on the other hand, spoken document retrieval (SDR) [4, 6] revolves more around the notion of relevance of a spoken document in response to the query. It is generally agreed upon that a document is relevant if it could address the stated information need of the query, not because it just happens to contain all the words in the given query [7, 8].

More recently, deep learning has gained significant interest of research and experimentation in many applications because of its remarkable performance [9, 10, 11]. When it comes to the field of natural language processing (NLP), word embedding methods can be viewed as pioneering studies [12-15]. A common thread of leveraging word embedding methods to NLP-related tasks is to represent a given paragraph (or sentence, document, and query) by simply taking an average over the word embeddings corresponding to the words occurring in the paragraph [16, 17]. As such, the similarity degree between a pair of paragraphs can be readily quantified by using one of the existing ranking mechanisms based on the learned representations. Celebrated methods developed in this vein include the continuous bag-of-words model [14], the skip-gram model [14, 18], the global vector model [15], to name just a few. Orthogonal to the NLP community, many research efforts have been devoted to the generative models because of the exceeding performance gains on image generation [19] and style transfer [20] tasks. The research trend can date back to the generative adversarial network (GAN) [21], which is mainly consisted of a generator and a discriminator. The discriminator, in general, is learned to distinguish the artificial samples from real distributions; the training objective for generator is to create high-quality and realistic samples, which expect to fool the trained discriminator. Since the two processes aim at optimizing their individual, but mutually conflicting, targets, a competitive minimax objective function is thus derived [22-24]. On top of the modeling principle, various GAN-based methods have demonstrated successful experiences and charming results in many applications [19, 20, 25].

One critical and practical issue facing SDR is that the input text/spoken query is usually too short to carry the information need of a user. In order to mitigate the fundamental challenge, a promising strategy is to reformulate the original query representation with extra statistics so as to boost the retrieval performance [26, 27]. The query reformulation methods devised following the line of research can be grouped into two distinct classes. One is to leverage external resources, such as Wikipedia or WordNet, to expand and reorganize the original query. The other is to reformulate the original query by referring to a small set of feedback documents locally collected from an initial round of retrieval, i.e., the so-called pseudo-relevance feedback (PRF) process [27, 28]. Since the former requires more sophisticated natural language processing techniques, including semantic representation and inference, as well as natural language generation, most efforts have been concentrated on launching the query reformulation methods by using the top-ranked feedback documents locally obtained from PRF [7, 29]. Although several studies have confirmed the effectiveness of PRF, the time-consuming problem makes it unappealing for realistic applications.

Motivated by the above observations, this paper strives to develop an efficient and effective modeling framework, which targets at generating pseudo-relevant statistics for a given query

automatically so as to re-estimate an enhanced query representation without performing the time-consuming PRF process. To sum up, the major contributions of this paper are at least three-fold. First, a novel framework, which not only concentrates on generating pseudo-relevant representations for a given query, but aims at excluding the limitation of time-wasting problem, is proposed. Second, stemming from such a framework, we thus propose an effective query reformulation method so as to enrich the original query. Finally, a series of empirical evaluations and comparisons are conducted on a benchmark SDR corpus.

2. RELATED WORK

2.1. The Classic Word Embedding Methods

The neural network language model [13] is the most-known seminal study on developing various word embedding methods. It estimates a statistical (*N*-gram) language model, formalized as a feed-forward fully-connected neural network, for predicting future words while inducing word embeddings as a by-product. Such an attempt has already motivated many follow-up extensions to develop similar methods for probing latent semantic and syntactic regularities in the representation of words. Representative methods include, but are not limited to, continuous bag-of-words (CBOW) model [14], skipgram (SG) model [14, 18], and global vector (GloVe) model [15].

Rather than seeking to learn a statistical language model, the CBOW, the SG, and the GloVe models manage to obtain a dense vector representation (embedding) of each word directly. The structure of CBOW is similar to a feed-forward fully-connected neural network, with the exception that the non-linear hidden layer in the former is removed. Formally, given a sequence of words, $w^1, w^2, \dots, w^T$, the objective function of CBOW is to maximize the log-probability for each segment of words [14]:

$$\sum_{t=1}^{T} \log \frac{\exp(\mathbf{v}_{\bar{w}^t} \cdot \mathbf{v}_{w^t})}{\sum_{i=1}^{|V|} \exp(\mathbf{v}_{\bar{w}^t} \cdot \mathbf{v}_{w_i})} \tag{1}$$

where $\mathbf{v}_{w^t}$ denotes the vector representation of the t^{th} word w^t in the training corpus, $\mathbf{v}_{\bar{w}^t}$ denotes the (weighted) average of vector representations of the contextual words of w^t, T denotes the length of the training corpus, and $|V|$ is the size of the vocabulary V. The concept of CBOW is motivated by the distributional hypothesis [30], which states that words with similar meanings often occur in similar contexts, and it is thus suggested to look for w^t whose word representation can capture its context distributions well. In contrast to the CBOW model, the SG model employs an inverse training objective with a simplified feed-forward fully-connected neural network [18]:

$$\sum_{t=1}^{T} \sum_{j=-c \& j \neq 0}^{c} \log \frac{\exp(\mathbf{v}_{w^{t+j}} \cdot \mathbf{v}_{w^t})}{\sum_{i=1}^{|V|} \exp(\mathbf{v}_{w_i} \cdot \mathbf{v}_{w^t})} \tag{2}$$

where c is the windows size of the contextual words for the central word w^t. Despite CBOW and SG models, the GloVe model suggests that an appropriate starting point for word representation learning should be associated with the ratios of co-occurrence probabilities rather than the prediction probabilities [15]. More precisely, GloVe makes use of weighted least squares regression, which aims at learning word representations by preserving the co-occurrence frequencies between each pair of words:

$$\sum_{i=1}^{|V|} \sum_{j=1}^{|V|} f\left(X_{w_i w_j}\right) \left(\mathbf{v}_{w_i} \cdot \mathbf{v}_{w_j} + b_{w_i} + b_{w_j} - \log\left(X_{w_i w_j}\right)\right)^2 \tag{3}$$

where $X_{w_i w_j}$ denotes the number of times words w_i and w_j co-occur in a pre-defined sliding context window; $f(\cdot)$ is a monotonic smoothing function used to modulate the impact of each pair of words involved in the model training; and $\mathbf{v}_w$ and b_w denote the word representation and the bias term of word w, respectively.

2.2. The Generative Adversarial Networks

In recent years, a popular research subject in the deep learning community is the generative adversarial networks (GANs) [21]. Opposite to classic research on neural networks, which mainly focuses on making decisions or regressions, GANs intend to build a generative model by neural networks. Generally, a GAN is at least comprised of two indispensable components, namely the generator $G(\cdot)$ and the discriminator $D(\cdot)$. The former tries to generate sharp and realistic samples, which close to the real data distribution, from a given noisy distribution, while the latter, on the contrary, concentrates on differentiating between real data and synthesized samples. Accordingly, the two competitive adversaries can be optimized by a minimax objective [21, 22]:

$$\begin{aligned} \min_G \max_D V(D,G) = {} & \mathbb{E}_{x \sim p_{data}(x)}\big(\log D(x)\big) \\ & + \mathbb{E}_{z \sim p_z(z)}\Big(\log\big(1 - D\big(G(z)\big)\big)\Big) \end{aligned} \tag{4}$$

where $p_{data}(\cdot)$ stands for real data distribution, and $p_z(\cdot)$ is a prior used to govern the input noisy data z. In practiced, the training procedure is performed iteratively by dictating that each network achieves optimally with the assumption that the other network is optimal.

3. LEARNING PSEUDO-RELEVANT REPRESENTATIONS

3.1 The Proposed Methodology

Due to the fact that a query usually consists of only a few words, the query representation of the query Q might not be accurately presented. With the alleviation of this deficiency as motivation, there are several studies dedicate to estimating a more accurate query representation, saying that it can be approached through a pseudo-relevance feedback (PRF) process [8, 27, 28]. The most simple but efficient method is the Rocchio's algorithm [7, 31], which introduces a mechanism of incorporating pseudo relevance feedback information into the vector space model [32]. Formally, for a given query Q, a set of top-ranked feedback documents $R_Q = \{d_1^Q, \cdots, d_{|R_Q|}^Q\}$ can be obtained by performing an initial-round of retrieval. After that, each feedback document (and the query) is expressed by a fixed-dimensional vector representation, which can be either composed by term frequency-inverse document frequency (TF-IDF) statistics or by taking an average over the word embeddings corresponding to the words occurring in the document (and query). Consequently, the Rocchio's algorithm adjusts the original query representation toward the center of these feedback document statistics [31]:

$$\mathbf{v}'_Q = \mathbf{v}_Q + \frac{1}{|R_Q|} \sum_{r=1}^{|R_Q|} \mathbf{v}_{d_r^Q} \tag{5}$$

where $\mathbf{v}_Q$ and $\mathbf{v}_{d_r^Q}$ denote the vector representations for query Q and document d_r^Q, respectively. Finally, the new query

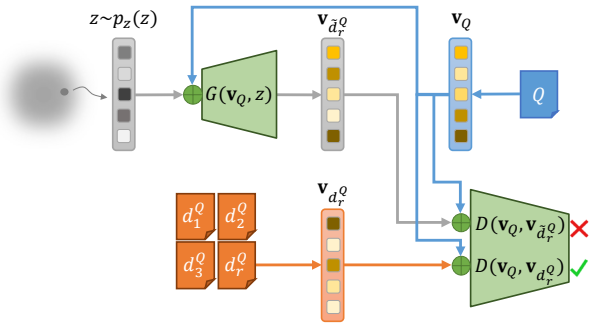

Figure 1. *Illustrations of the proposed framework.*

representation $\mathbf{v}_Q'$ can be used to select relevant documents based on the cosine similarity measure. Although several query reformulation methods based on PRF have been proposed and proven their success in various IR-related tasks, the time-consuming problem makes it unappealing for realistic applications. To overcome this deficiency, we hence investigate a novel framework, which aims at generating a set of pseudo-relevant representations automatically and also deducing a more informative and robust vector representation for a user's query so as to boost the retrieval performance.

To crystallize the idea to go, we begin the framework by employing a generator $G(\cdot)$ and a discriminator $D(\cdot)$. The former is used to synthesize a set of pseudo-relevant representations for a given query, and the latter is introduced to criticize the generator as well as guide the model training. More formally, for a training query Q and its corresponding feedback documents $R_Q = \{d_1^Q, \cdots, d_{|R_Q|}^Q\}$, the query and documents are first represented by vector representations, where the vector is simply taking an average over the word embeddings corresponding to the words occurring in the query/document. Then, the generator learns a mapping from the observed query representation $\mathbf{v}_Q$ and a random noisy vector z to a generated representation $G(\mathbf{v}_Q, z)$. Since the ultimate goal of $G(\cdot)$ is to synthesize a set of representations, which are not only relevant but complementary to the given query, the discriminator is thus introduced. The discriminator is trained to do as well as possible at detecting the generator's creations. Accordingly, $D(\cdot)$ takes the original query representation $\mathbf{v}_Q$, and a feedback document representation (i.e., $\mathbf{v}_{d_r^Q}$) or a synthesized representation (i.e., $G(\mathbf{v}_Q, z)$) as input. After that, a decision score is obtained to indicate whether the input is real (i.e., $\mathbf{v}_{d_r^Q}$) or fake (i.e., $G(\mathbf{v}_Q, z)$). Hence, the discriminator not only has to distinguish fake examples (i.e., $G(\mathbf{v}_Q, z)$) from true ones (i.e., $\mathbf{v}_{d_r^Q}$), but also needs to quantify the relationship between query and document representations. Consequently, a minimax objective function is derived to optimize the two adversaries:

$$\min_G \max_D V(D,G) = \mathbb{E}_{\mathbf{v}_{d_r^Q} \sim p_Q(\mathbf{v}_{d_r^Q})} \left(\log D\left(\mathbf{v}_Q, \mathbf{v}_{d_r^Q}\right)\right) + \mathbb{E}_{z \sim p_z(z)} \left(\log\left(1 - D\left(\mathbf{v}_Q, G(\mathbf{v}_Q, z)\right)\right)\right) \quad (6)$$

where $p_Q(\cdot)$ stands for real data distribution, and $p_z(\cdot)$ is a prior used to govern the input noisy data z. By doing so, the generator anticipates to generate a set of representations, which act as those pseudo-relevant documents selected by PRF. In other words, the generator can produce fantastically realistic pseudo-relevant representations, which can fool the discriminator, for a given query. It is worthy to note that the proposed mechanism is similar to the condition-based GAN methods [33], but the major difference is that our method can be viewed by conditioning on a continuous random variable (i.e., $\mathbf{v}_Q$), while the classic methods usually condition on a well-defined and finite discrete category. Figure 1 schematically depicts the architecture of the proposed framework.

3.2 The Retrieval Model

Subsequently, based on a set of generated pseudo-relevant representations, denoted by $\tilde{R}_Q = \{\mathbf{v}_{\tilde{d}_1^Q}, \cdots, \mathbf{v}_{\tilde{d}_{|\tilde{R}_Q|}^Q}\}$, for a given query Q, we turn to reformulate the original query so as to boost the SDR performance. It is usually anticipated that the SDR system can thus probe more relevant documents by the enhanced query. Inspired by the Rocchio's algorithm [31], an intuitive and straightforward strategy is to pool every generated representation weighted by its discriminant score, which distinguishes highly relevant representations from less relevant ones, to yield a new query representation:

$$\mathbf{v}_Q' = \beta \cdot \mathbf{v}_Q + (1-\beta) \cdot \left(\sum_{r=1}^{|\tilde{R}_Q|} d(\mathbf{v}_Q, \mathbf{v}_{\tilde{d}_r^Q}) \cdot \mathbf{v}_{\tilde{d}_r^Q} \right) \quad (9)$$

where $d(\mathbf{v}_Q, \mathbf{v}_{\tilde{d}_r^Q})$ is the normalized discriminant score for each representation $\mathbf{v}_{\tilde{d}_r^Q}$, and β is a weighting factor to modulate the balance between the original query and the generated information. The discriminant score for $\mathbf{v}_{\tilde{d}_r^Q}$ is given by the trained discriminator, that is $D(\mathbf{v}_Q, \mathbf{v}_{\tilde{d}_r^Q})$, in this sudy. With the purpose of blending both literal and semantic information together, the proposed framework is thus linearly combined with the classic VSM, which determines the relevance score by referring to term frequency-inverse document frequency (TF-IDF) statistics for both query and document. We term the entire process a generating pseudo-relevant representations framework and denote hereafter as GPR in short. To sum up, in the retrieval/test stage, the GPR will first generate a set of pseudo-relevant representations for a given query. After that, an enhanced query representation will be constructed by using these synthesized representations. Finally, the enhanced query vector is used to rank the documents. Therefore, it is worth mentioning that the GPR can estimate an enhanced query representation without performing the time-consuming PRF process, and we anticipate to obtain a better retrieval result by referring to the new query vector.

4. EXPERIMENTS

4.1 Experimental Setup

We used the Topic Detection and Tracking collection (TDT-2) [34] in the experiments. The Mandarin news stories from Voice of America news broadcasts were used as the spoken documents. All news stories were exhaustively tagged with event-based topic labels, which served as the relevance judgments for performance evaluation. The average word error rate obtained for the spoken documents is about 35% [35]. The Chinese news stories from Xinhua News Agency were used as our test queries. More specifically, in the following experiments, we will either use a whole news story as a "long query," or merely extract the tittle field from a news story as a "short query." The retrieval performance is evaluated with the commonly-used non-interpolated mean average precision (MAP) following the TREC evaluation [8]. In this study, both the generator $G(\cdot)$ and the discriminator $D(\cdot)$ are implemented by fully

connected deep networks with different model parameters θ_D and θ_G, the optimizer is stochastic gradient decent method (SGD), and the activation function used in both $G(\cdot)$ and $D(\cdot)$ is the hyperbolic tangent, except that the output layer in the discriminator adopts the sigmoid. To obtain the model parameters, 819 training query exemplars with the corresponding top-ranked feedback documents are compiled. Based on that, the training instance is generated by 1) randomly selecting a training query, 2) picking one of its feedback documents to the query to be a *true* example, 3) randomly choosing a random vector from the noisy distribution, and 4) generating a synthesized representation to form a *fake* example.

4.2 Experimental Results

In the first set of experiments, we explore the efficacies of several baseline systems, including the vector space model (VSM) [32], three classic word embedding methods [14, 15] (i.e., CBOW, SG, and GloVe), two classic paragraph embedding methods (i.e., DM [36] and EV [37]), and the classic Rocchio's algorithm [31] for SDR. For all the baseline systems, each query and document is represented by a vector, and the relevance degree is computed by the cosine similarity measure. Furthermore, in this study, we also make a comparison between SDR and traditional text retrieval. Consequently, the retrieval results, assuming manual transcripts for the spoken documents to be retrieved (denoted by TD) are known, are also shown for reference, compared to the results when only the erroneous transcripts by speech recognition are available (denoted by SD). Experimental results are shown in the first block of Table 1. The best result within each column (corresponding to a specific evaluation condition) is type-set boldface. Inspection of these results reveals five noteworthy points. First, all of the word embedding and paragraph embedding-based methods outperform VSM by a large margin. Second, SG demonstrates superior results over CBOW and GloVe, which is consistent with other studies on several tasks. Third, for paragraph embedding-based methods, EV, which learns to distill the most important information from the paragraph and exclude general background information, appears to be more flexible than DM, thereby yielding better results. Fourth, the Rocchio's algorithm outperforms all the other methods in most cases, which indicates that the PRF process can really benefit the performance of retrieval. Finally, the performance gap between the retrieval on the manual transcripts (i.e., the TD case) and that on the recognition transcripts (i.e., the SD case) is about 6% in terms of MAP, which also shows that the recognition errors inevitably mislead the statistics for both query and document so as to degrade the retrieval performance.

Next, we start to evaluate the proposed GPR framework for SDR. The results are presented in the second block of Table 1. At the first glance, the empirical results reveal that GPR achieves better results than various baseline systems in most cases. It is worthy to note that the Rocchio's algorithm can be treated as our major challenger, because the Rocchio's algorithm leverages the time-consuming PRF process to collect a set of feedback documents so as to reformulate a new query representation, while the proposed GPR infers an enhanced query vector by using a set of automatically synthesized representations. Consequently, we are supervised that GPR achieves comparable (or even better) results than Rocchio's algorithm, which may implicitly imply that the generated representations are very robust and even can be used to replace the real ones. Furthermore,

Table 1. *Retrieval results (in MAP) achieved by various retrieval systems.*

	TD		SD	
	Long	Short	Long	Short
VSM	0.548	0.339	0.484	0.273
CBOW	0.563	0.358	0.500	0.307
SG	0.567	0.385	0.508	0.364
GloVe	0.558	0.371	0.502	0.321
DM	0.558	0.344	0.484	0.302
EV	0.571	0.382	0.518	0.364
Rocchio's	**0.577**	**0.385**	**0.526**	**0.389**
GPR	0.584	0.404	0.523	0.380
GPR+Rocchio's	**0.589**	**0.404**	**0.527**	**0.389**

Table 2. *Retrieval results (in MAP) of the proposed GPR framework with respect to the number of generated representations.*

	TD		SD	
$\lvert\tilde{R}_Q\rvert$	Long	Short	Long	Short
1	0.573	0.386	0.509	0.370
3	**0.584**	0.387	0.508	0.364
5	0.576	**0.404**	**0.523**	0.363
10	0.571	0.386	0.510	**0.380**

we make a step forward to linearly combine GPR and Rocchio's algorithm (denoted by GPR+Rocchio's). The results are also summarized in Table 1. As expected, the combination only achieves a small performance gain when compared to Rocchio's algorithm or the proposed GPR, respectively. Thus, we can indeed conclude that the synthesized representations are not only robust but also really similar to those "real" representations. In the last set of experiments, we look into the impact of the number of generated documents on the GPR framework. As revealed by the results illustrated in Table 2, leveraging a small number of representations (e.g., 3 and 5) seems to be adequate for the TD cases, while a large number of representations (e.g., 5 and 10) seems to be suitable for the SD cases. This can be attributed to the fact that more extra statistics seems to be more robust to the recognition errors. Nevertheless, the way to systemically determine the optimal number of representations remains an open issue and needs further investigation. To sum up, a series of empirical experiments demonstrate the good efficacy and capacity of the proposed GPR framework.

5. CONCLUSION

In this paper, we have presented a novel framework, which can be leveraged to generate a set of pseudo-relevant representations for a given query. The framework has been evaluated on a SDR benchmark corpus. Experimental results demonstrate the remarkable superiority than other strong baselines compared in the paper, thereby indicating the potential of the new GPR framework. For future work, we will explore the incorporation of extra cues, such as acoustic statistics and sub-word information, into the proposed framework for the SDR task. Moreover, we also plan to evaluate the framework on other large-scale corpora as well as NLP-related tasks.

6. ACKNOWLEDGMENT

This work was supported in part by the Ministry of Science and Technology of Taiwan under Grants MOST 106-2218-E-011-019-MY3 and MOST 108-2636-E-011-005, as well as by the Chunghwa Telecom Laboratories under Grant TL-107-8201.

7. REFERENCES

[1] L.-S. Lee and B. Chen, "Spoken document understanding and organization," *IEEE Signal Processing Magazine*, 22(5):42–60, 2005.

[2] C. Chelba, T. J. Hazen and M. Saraclar, "Retrieval and browsing of spoken content," *IEEE Signal Processing Magazine*, 25(3), 2008.

[3] M. Ostendorf, "Speech technology and information access," *IEEE Signal Processing Magazine*, pp. 150–152, 2008.

[4] M. Larson and G. J. F. Jones, "Spoken content retrieval: a survey of techniques and technologies", *Foundations and Trends in Information Retrieval*, 5(4–5):235–422, 2012.

[5] I. Szoke, L. Burget, J. Cernocky and M. Fapso, "Sub-word modeling of out of vocabulary words in spoken term detection," in *Proc. of SLT*, 2008.

[6] C. Carpineto and G. Romano, "A survey of automatic query expansion in information retrieval," *ACM Comput. Surv.*, 44(1), 2012.

[7] C. D. Manning, P. Raghavan and H. Schtze, *Introduction to information retrieval*, New York: Cambridge University Press, 2008.

[8] R. A. Baeza-Yates and B. Ribeiro-Neto, *Modern information retrieval: the concepts and technology behind search*, Addison-Wesley Longman Publishing Co., Inc., 2011.

[9] Y. LeCun, Y. Bengio and G. Hinton, "Deep learning," *Nature*, 521:436–444, 2015.

[10] J. Schmidhuber, "Deep learning in neural networks: an overview," *Neural Networks*, 61:85–117, 2015.

[11] I. Goodfellow and Y. Bengio and A. Courville, *Deep learning*, MIT Press, 2016.

[12] T. Young, D. Hazarika, S. Poria, E. Cambria "Recent trends in deep learning based natural language processing," arXiv:1708.02709, 2017.

[13] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A neural probabilistic language model," *Journal of Machine Learning Research* (3), pp. 1137–1155, 2003.

[14] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proc. of ICLR*, 2013.

[15] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vector for word representation," in *Proc. of EMNLP*, 2014.

[16] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proc. of ICML*, pp. 1188–1196, 2014.

[17] K.-Y. Chen, S.-H. Liu, H.-M. Wang, B. Chen, and H.-H. Chen "Leveraging word embeddings for spoken document summarization," in *Proc. of INTERSPEECH*, 2015.

[18] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. of ICLR*, 2013.

[19] K. Gregor, I. Danihelka, A. Graves, D. J. Rezende, and D. Wierstra, "DRAW: A Recurrent Neural Network For Image Generation," arXiv:1502.04623*,* 2015.

[20] L. A. Gatys, A. S. Ecker, and M. Bethge, "A neural algorithm of artistic style", arXiv:1508.06576, 2015.

[21] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y Bengio, "Generative adversarial networks," in *Proc. of NIPS*, 2014.

[22] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *Proc. of NIPS*, 2016.

[23] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. of ICML*, 2017.

[24] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," in *Proc. of ICLR*, 2018.

[25] L. Yu, W. Zhang, J. Wang, Y. Yu, "SeqGAN: Sequence generative adversarial nets with policy gradient," in *Proc. of AAAI*, 2018.

[26] J. Huang and E. N. Efthimiadis, "Analyzing and evaluating query reformulation strategies in web search logs," in *Proc. of CIKM*, 2009.

[27] K.-Y. Chen, S.-H. Liu, B. Chen, E.-E. Jan, H.-M. Wang, W.-L. Hsu, and H.-H. Chen, "Leveraging effective query modeling techniques for speech recognition and summarization," in *Proc. of EMNLP*, 2014.

[28] S. Clinchant and E. Gaussier, "A theoretical analysis of pseudo-relevance feedback models," in *Proc. of ICTIR*, 2013.

[29] Y. Lv and C.-X. Zhai, "A comparative study of methods for estimating query language models with pseudo feedback," in *Proc. of CIKM*, 2009.

[30] G. Miller and W. Charles, "Contextual correlates of semantic similarity," *Language and Cognitive Processes*, 6(1):1–28, 1991.

[31] J. J. Rocchio, "Relevance Feedback in information retrieval," in *Proc. of SIGIR*, 1965.

[32] G. Salton , A. Wong , C. S. Yang, "A Vector Space Model for automatic indexing," *Communications of the ACM*, pp. 613–620, 1975.

[33] M. Mirza and S. Osindero, "Conditional generative adversarial nets," arXiv:1411.1784, 2014.

[34] LDC, "Project topic detection and tracking," *Linguistic Data Consortium*, 2000.

[35] H. Meng, S. Khudanpur, G. Levow, D. Oard, and H.-M. Wang, "Mandarin–English information (MEI): investigating translingual speech retrieval," *Computer Speech and Language*, 18(2):163–179, 2004.

[36] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proc. of ICML*, 2014.

[37] K.-Y. Chen, S.-H. Liu, B. Chen, H.-M. Wang, "Essense vector-based query modeling for spoken document retrieval," in *Proc. of ICASSP*, 2018.