INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Exploring Word Mover's Distance and Semantic-Aware Embedding Techniques for Extractive Broadcast News Summarization

Shih-Hung Liu[1, 2], Kuan-Yu Chen[1, 2], Yu-Lun Hsieh[1], Berlin Chen[3], Hsin-Min Wang[1], Hsu-Chun Yen[2], Wen-Lian Hsu[1]

[1] Academia Sinica, Taiwan
[2] National Taiwan University, Taiwan
[3] National Taiwan Normal University, Taiwan

[1]{journey,kychen,morphe,whm,hsu}@iis.sinica.edu.tw,
[2] yen@cc.ee.ntu.edu.tw, [3] berlin@ntnu.edu.tw

Abstract

Extractive summarization is a process that manages to select the most salient sentences from a document (or a set of documents) and subsequently assemble them to form an informative summary, facilitating users to browse and assimilate the main theme of the document efficiently. Our work in this paper continues this general line of research and its main contributions are two-fold. First, we explore to leverage the recently proposed word mover's distance (WMD) metric, in conjunction with semantic-aware continuous space representations of words, to authentically capture finer-grained sentence-to-document and/or sentence-to-sentence semantic relatedness for effective use in the summarization process. Second, we investigate to combine our proposed approach with several state-of-the-art summarization methods, which originally adopted the conventional term-overlap or bag-of-words (BOW) approaches for similarity calculation. A series of experiments conducted on a typical broadcast news summarization task seem to suggest the performance merits of our proposed approach, in comparison to the mainstream methods.

Index Terms: extractive summarization, word representation, word mover's distance, Markov random walk

1. Introduction

With the emergence of the big data era, unlimited amounts of multimedia such as TED talks, broadcast news programs, online video sharing websites, etc., has overwhelmed our daily life [1-4]. This inevitably leads to the information overload problem. Viable summarization techniques are is highly in demand to alleviate the problem and enable people to efficiently browse or digest the multimedia content by either listening or reading. Automated extractive summarization bears the goal of producing compact summary sentences from the source document according to a preferred ratio. Among major aspects that should be considered in the summarization process, relevance is of paramount importance for characterizing sentence-to-document and/or sentence-to-sentence relationships. Basically, extractive summarization can be framed as a ranking process that extracts the most salient set of sentences to form an informative and concise summary.

Existing methods for extractive speech summarization developed so far can be roughly divided into three categories, including those based on the sentence position or structure information, unsupervised sentence ranking, and those relying on supervised sentence classification [5-8]. A common practice for most of the unsupervised methods is to select important sentences by means of some statistical features of sentences or of the words in the sentences, where feature extraction and model estimation are typically conducted without human supervision. Statistical features, for example, can be the term (word) frequency, linguistic score and recognition confidence measure, as well as prosodic information. On top of these indicative features, numerous representative methods have been introduced, including the vector space model (VSM) [9], latent semantic analysis (LSA) [9], Markov random walk (MRW) [11], maximum marginal relevance (MMR) [10], LexRank [12], submodularity-based method (SM) [13], integer linear programming (ILP) [14] and language modeling approach [15-17], among others.

Recently, in the natural language processing (NLP) community, representation learning for words has become an active research topic [18, 19] and spurred a vast range of downstream applications. The essence of these methods is to learn continuously distributed (as opposed to one-hot) vector representations of words using neural networks. Accordingly, the learned representations can encode latent semantic and/or syntactic information and in turn can be used to infer similarity/relevance among words. A common thread of leveraging word embedding methods in NLP-related tasks is to represent a portion of text (e.g., paragraph, sentence, or document) by averaging the corresponding word embeddings over all words in the desired portion. After that, the cosine similarity measure, as a straightforward choice, can be readily applied to determine the degree of relevance between any pair of representations.

Although the utilities and abilities of representing a portion of text by taking average of the word representations within it have been proven recently, the composite representation may drift the main theme of the semantic content or could not accentuate those indicative words in the piece of text. As a result, the determined relevance degree between a pair of representations might be undesired. Beyond the continued and tremendous efforts made to develop the representation methods for words, this paper focuses on mitigating the fundamental downside in recent studies and propose two major contributions. On one hand, the word mover's distance (WMD) metric [25] that builds on top of the word embedding

http://dx.doi.org/10.21437/Interspeech.2016-710

space recently has been proposed to accurately estimate the similarity degree between a pair of documents. However, as far as we are aware, this notion has never been extensively explored in extractive text or speech summarization. The paper sets out to leverage the WMD metric, in conjunction with semantic-aware continuous space representations of words, to authentically capture finer-grained sentence-to-document and/or sentence-to-sentence semantic relationships for effective use in the extractive speech summarization. On the other hand, we also investigate to combine WMD with several state-of-the-art summarization methods, which originally adopted the conventional term-overlap or bag-of-words (BOW) approaches for similarity calculation.

2. Previous Work

2.1. Continuous Word Representation Methods

One of the most-known pioneering studies on developing word embedding methods was presented in [19]. It estimated a statistical (*n*-gram) language model, formalized as a feed-forward neural network, for predicting future words in context while inducing word embeddings (or representations) 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 a word. Representative methods include, but are not limited to, the continuous bag-of-words (CBOW) model [18, 19] and the skip-gram (SG) model [19, 20].

Rather than seeking to learn a statistical language model, the CBOW model manages to obtain a dense vector representation (embedding) of each word directly. The structure of CBOW is similar to a feed-forward neural network, with the exception that the non-linear hidden layer in the former is removed. By getting around the heavy computational burden incurred by the non-linear hidden layer, the model can be trained on a large corpus efficiently, while still retains good performance. Formally, given a sequence of words, w_1, w_2… w_T, the objective function of CBOW is to maximize the log-probability

$$\sum_{t=1}^{T} \log P(w_t \mid w_{t-c},...,w_{t-1},w_{t+1},...,w_{t+c}), \tag{1}$$

where c is the window size of the contextual words for the central word w_t and T denotes the length of the training corpus.

In contrast to the CBOW model, the SG model employs an inverse training objective for learning word representations with a simplified feed-forward neural network [19-21]. Given a sequence of words, w_1, w_2… w_T, the objective function of SG is to maximize the log-probability

$$\sum_{t=1}^{T} \sum_{j \in neighbor(t)} \log P(w_j \mid w_t), \tag{2}$$

where *neighbor*(t) is the set of neighboring words of word w_t. The concept of the above two word embedding methods is motivated by the distributional hypothesis, 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 the distributions of its context well.

2.2. Graph-based Summarization Methods

The graph-based summarization methods conceptualize the document to be summarized as a graph of sentences, where each node represents a sentence and the associated weight of each link represents the lexical similarity relationship between a pair of nodes. The Markov random walk (MRW) and the LexRank [12] are two representatives. Text or speech summarization thus relies on the global structural information embedded in such a sentence graph, rather than merely considering the similarity solely between each sentence of the document to be summarized and the document itself. Put simply, sentences that are more similar to others are deemed more salient to the main theme of the document [11].

Formally, taking MRW as an example, an affinity matrix (or a graph) that represents the relationships between sentences is firstly constructed in response to a document to be summarized. The relevance degree between a pair of sentences (i.e., an entry in the matrix or the weight of an edge in the graph) is determined by calculating the cosine similarity measure. Afterwards, a normalization process is performed row by row of the affinity matrix to result in a stochastic matrix *M*. Based on the stochastic matrix *M*, the MRW-based algorithm [11] is recursively performed to obtain the saliency score $R(S_i)$ for each sentence S_i:

$$R(S_i) = \mu \cdot \sum_{all\ j \neq i} R(S_j) \cdot M_{ji} + \frac{(1-\mu)}{|D|}, \tag{3}$$

where μ is a damping factor in order to achieve aperiodicity of this iterative Markov process, $|D|$ is the number of sentences in the given document, and M_{ji} denotes the transition probability from sentence S_j to sentence S_i.

On the other hand, LexRank bears a close resemblance to MRW by selecting salient sentences based on the notion of eigen-centrality of a sentence graph [12]. The major difference between LexRank and MRW is that the former uses the degree of a node (sentence) to build the stochastic matrix while the latter creates the affinity matrix by the cosine similarity measure. Both MRW and LexRank in essence are inspired from the well-known PageRank algorithm [22] that is widely adopted by most of today's commercial search engines on the Internet.

3. Word Mover's Distance

3.1. The Principle

Instead of determining the similarity degree between a pair of sentences (or documents) simply with the composite representations of words, the principle of word mover's distance (WMD) [25] assumes that the dissimilarity score between two words is a natural building block to measure the distance between two sentences.

Building on the assumption, WMD first defines the individual distance (or travel cost) between a pair of words w_i and w_j corresponding to their learned word embeddings $e(w_i)$ and $e(w_j)$:

$$d(w_i, w_j) = \| e(w_i) - e(w_j) \|_2, \tag{4}$$

Accordingly, WMD formulates the dissimilarity degree between a pair of sentences, S and S', by calculating the minimum amount of summing up individual distances (travel costs) that words occurring in S need to travel to reach the words occurring in S':

$$d_{\mathrm{WMD}}(S,S') = \min_{F \geq 0} \sum_{w_i \in S} \sum_{w_j \in S'} F_{w_i w_j} \times d(w_i, w_j),$$
$$s.t. \sum_{w_j \in S'} F_{w_i w_j} = \frac{c(w_i, S)}{|S|}, \forall w_i \in S \quad (5)$$
$$\sum_{w_i \in S} F_{w_i w_j} = \frac{c(w_j, S')}{|S'|}, \forall w_j \in S'$$

where $F \in \mathbb{R}^{V \times V}$ is a flow matrix which indicates how much probability mass should flow (or travel) from word w_i in sentence S to word w_j in another sentence S', and vice versa. Theoretically, we can allow each word to be transformed into any word in total or in parts. Furthermore, the first constraint denotes that the entire outgoing flow from word w_i equals to its own probability mass in sentence S; meanwhile, the second constraint signals that the entire incoming flow from word w_j can only equal to its own probability mass in S'. The optimization problem is a special case of the earth mover's distance metric (EMD) [25, 26], a well-known transportation problem, and there exist some specialized solvers that can be readily applied to this problem [27].

3.2. The Proposed Summarization Framework

Thanks to the word embedding techniques and the WMD metric, a semantic-aware dissimilarity measure can thus be obtained in a systematic and theoretically sound manner. Here we make use of the WMD measure in the summarization process by integrating it into the Markov random walk algorithm. For the idea to go, we begin by converting the WMD measure between any pair of sentences in a given document to be summarized to a similarity score:

$$sim(S,S') = \frac{1}{1 + e^{\alpha(d_{\mathrm{WMD}}(S,S') - avg(D))}}, \quad (6)$$

where α is a tunable parameter used to control the slope of the sigmoid function; $d_{WMD}(S, S')$ denotes the WMD measure between sentence S and sentence S'; and $avg(D)$ is the document-specific parameter which is computed by taking average of all the WMD measures for all pairs of sentences involved in the document D. A pair of sentences that has a smaller WMD distance will have a higher similarity, and vice versa. By converting the dissimilarity score to a similarity score through the sigmoid function, we can build the affinity matrix M, which is used to replace the original affinity matrix (*cf.* M in Eq. (3)) constructed with the conventional cosine similarity measure, and the enhanced summarization method can naturally select salient sentences to form a concise summary for a given document.

4. Experiments

4.1. Experiment Setup

We use the MATBN broadcast news corpus collected by the Academia Sinica and the Public Television Service Foundation of Taiwan between November 2001 and April 2003 [29]. It has been segmented into separate stories and transcribed manually. Each story contains the speech of one news anchor, as well as several field reporters and interviewees. A subset of 205 broadcast news documents between November 2001 and August 2002 was reserved for the summarization experiments. We chose 20 documents as the test set, while the remaining 185 documents as the development set. The reference summaries were generated by ranking the sentences in the manual transcript of a spoken document by importance without assigning a score to each one. Each document has three reference summaries annotated by three human subjects. For the assessment of summarization performance, we adopted the widely-used ROUGE metrics [30]. All experimental results reported hereafter are obtained by calculating the F-scores [30] of these ROUGE metrics. The summarization ratio was set to 10%. A subset of 25-hour speech data from MATBN compiled from November 2001 to December 2002 was used to bootstrap acoustic model training with a minimum phone error rate (MPE) criterion and a training data selection scheme [31]. The vocabulary size is about 72 thousand words. The average word error rate of automatic transcription is about 40%.

Finally, a larger corpus containing 100,000 text news documents collected by the Central News Agency (CNA) between 2000 and 2001 (the Chinese Gigaword Corpus released by LDC) during the same period as the broadcast news documents to be summarized, were used for training the CBOW and SG word embedding models.

4.2. Experimental Results

At the outset, we compare the performance of a naïve VSM baseline against two popular word embedding methods (i.e. CBOW and SG), as well as their pairing with WMD. The results are shown in Table 1, where TD denotes the results obtained based on the manual transcripts of spoken documents and SD denotes the results using the automatic speech recognition transcripts that may contain recognition errors. In Table 1, we observe that the two embedding methods, though with different model structures and learning strategies, achieve results comparable with each other in both the TD and SD cases. WMD_SG (denotes the WMD distance in the SG semantic vector space) outperforms WMD_CBOW in the TD case and offers moderate performance in SD case, though the difference is less pronounced. The results also indicate that WMD-based methods outperform the word embedding methods as expected in the TD case, whereas they perform worse than the word embedding methods in the SD case. The reason may be that the recognition errors affect the WMD methods more severely than the word embedding methods. Since WMD is calculated by considering each and every word in a spoken sentence, while the word embedding methods use an average word embedding. We postulate that the effect of averaging could dampen the negative impact of imperfect speech recognition.

In the next set of experiments, we evaluate the effectiveness of the graph-based ranking method (MRW) with various affinity matrices constructed by VSM (using term frequency weighted by inverse document frequency), CBOW and SG with the cosine measure, and WMD with CBOW and SG embeddings. The corresponding results are shown in Table 2. Comparing Tables 1 and 2, we can conclude that the MRW process indeed boosts the summarization performance for all

methods in both the TD and SD cases. A closer look at Table 2 reveals that WMD_SG is the best-performing method in the TD case, while this claim is reversed again for the SD case owing presumably to the recognition errors. Nevertheless, the performance of WMD_SG and WMD_CBOW in the TD case validates the utility of the similarity measure induced by the proposed WMD method, which operates on the well-constructed semantic space. A possible improvement for the WMD-based methods to cope with imperfect speech recognition is to leverage subword units for building subword-based semantic space (such as syllable embeddings). It is worthy of future investigation.

In the last set of experiments, we assess the performance levels of several well-practiced or/and state-of-the-art methods for extractive summarization, including position-based method (LEAD) [32], variants of the vector-space model (i.e., latent semantic analysis (LSA), maximum marginal relevance (MMR)), the language model-based summarization method (i.e., unigram language model (ULM)), the graph-based methods (i.e., Markov random walk (MRW) and LexRank), and combinatorial optimization methods (i.e., Submodularity (SM) and integer linear programming (ILP)). The results are depicted in Table 3, along with the best results obtained by the proposed WMD+MRW method coupling with SG embeddings. Several noteworthy observations can be drawn. First, LSA, which represents the sentences of a spoken document and the document itself in the latent semantic space instead of the index term (word) space, performs slightly better than VSM in both the TD and SD cases (*cf.* Table 1). The two graph-based methods (i.e., MRW and LexRank) are quite comparable with each other, and perform better than the vector-space methods (i.e., VSM, LSA, CBOW, and SG) in the TD case. However, in the SD case, the situation is reversed. It reveals that imperfect speech recognition seems to have a stronger negative influence on the graph-based ranking methods than the vector-space methods. This may be attributed to the speech recognition errors that lead to inaccurate calculation of the relevance measure between each pair of sentences. The PageRank-like procedure of the graph-based methods, in turn, will be executed based on these problematic measures, potentially leading to degraded results. Notably, ULM shows comparable results to the other state-of-the-art methods for both the TD and SD cases, demonstrating the strength of the language modeling approach. For the combinatorial methods (i.e., SM and ILP), they stand out in performance for the TD case, but only deliver results on par with the other methods in the SD case. Although both SM and ILP aptly integrate the ability of reducing redundancy (or increasing diversity) for summarization, they involve higher time complexity [14]. Last, the proposed WMD+MRW method that integrates WMD within MRW can achieve comparable results with the state-of-the-art methods for the TD case and outperform them for the SD case. These results again demonstrate the good potential of the proposed method.

Table 1. *Summarization results achieved by the word embedding methods, WMD and their combinations.*

Method	Text Documents (TD)			Spoken Documents (SD)		
	Rouge-1	Rouge-2	Rouge-L	Rouge-1	Rouge-2	Rouge-L
VSM	0.347	0.228	0.290	0.342	0.189	0.287
CBOW	0.382	0.249	0.322	0.362	0.214	0.314
SG	0.371	0.239	0.311	0.364	0.215	0.311
WMD_CBOW	0.384	0.258	0.329	0.331	0.169	0.281
WMD_SG	0.401	0.280	0.348	0.336	0.174	0.283

Table 2. *Summarization results achieved by incorporating the word embedding methods and WMD in the MRW process.*

Method	Text Documents (TD)			Spoken Documents (SD)		
	Rouge-1	Rouge-2	Rouge-L	Rouge-1	Rouge-2	Rouge-L
VSM	0.412	0.282	0.358	0.332	0.191	0.291
CBOW	0.436	0.310	0.384	0.393	0.246	0.346
SG	0.416	0.283	0.351	0.375	0.232	0.323
WMD_CBOW	0.432	0.312	0.372	0.361	0.199	0.300
WMD_SG	0.442	0.329	0.387	0.371	0.217	0.317

Table 3. *Summarization results achieved by the proposed MRW+WMD method and some state-of-the-art unsupervised methods.*

Method	Text Documents (TD)			Spoken Documents (SD)		
	Rouge-1	Rouge-2	Rouge-L	Rouge-1	Rouge-2	Rouge-L
LEAD	0.310	0.194	0.276	0.255	0.117	0.221
LSA	0.362	0.233	0.316	0.345	0.201	0.301
MMR	0.368	0.248	0.322	0.366	0.215	0.315
ULM	0.411	0.298	0.361	0.364	0.210	0.307
MRW	0.412	0.282	0.358	0.332	0.191	0.291
LexRank	0.413	0.309	0.363	0.305	0.146	0.254
SM	0.414	0.286	0.363	0.332	0.204	0.303
ILP	0.442	0.337	0.401	0.348	0.209	0.306
WMD+MRW	0.442	0.329	0.387	0.371	0.217	0.317

5. Conclusions and Future Work

In this paper, a novel word mover's distance-based methods standing on the solid ground of a semantic vector space representation technique have been proposed and evaluated for extractive spoken document summarization. Moreover, incorporating them into the start-of-the-art graph-based ranking process leads to better performance in the selection of indicative sentences. Experimental results confirm the effectiveness of the methods instantiated from our summarization framework, in comparison to several celebrated methods. For future work, we will explore other effective ways to enrich the representations of words and integrate additional cues, such as speaker identities or prosodic (emotional) information, into the proposed framework. We are also interested in investigating a more robust representation for spoken documents in order to offset the negative impact of imperfect speech recognition.

6. Acknowledgements

This research is supported in part by the "Aim for the Top University Project" of National Taiwan Normal University (NTNU), sponsored by the Ministry of Education, Taiwan, and by the Ministry of Science and Technology, Taiwan, under Grants MOST 103-2221-E-003-016-MY2, MOST 104-2221-E-003-018-MY3, MOST 104-2911-I-003-301.

7. References

[1] S. Furui *et al.*, "Fundamental technologies in modern speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 16–17, 2012.

[2] M. Ostendorf, "Speech technology and information access," *IEEE Signal Processing Magazine*, vol. 25, no. 3, pp. 150–152, 2008.

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

[4] L. S. Lee *et al.*, "Spoken content retrieval—beyond cascading speech recognition with text retrieval," IEEE/ACM Transactions on Audio, Speech and Language Processing, vol. 23, no. 9, pp. 1389–1420, 2015.

[5] Y. Liu and D. Hakkani-Tur, "*Speech summarization*," Chapter 13 in Spoken Language Understanding: Systems for Extracting Semantic Information from Speech, G. Tur and R. D. Mori (Eds), New York: Wiley, 2011.

[6] A. Nenkova and K. McKeown, "Automatic summarization," Foundations and Trends in Information Retrieval, vol. 5, no. 2–3, pp. 103–233, 2011.

[7] I. Mani and M. T. Maybury (Eds.), Advances in automatic text summarization, Cambridge, MA: MIT Press, 1999.

[8] J.-M. Torres-Moreno (Eds.), "Automatic text summarization," WILEY-ISTE, 2014.

[9] Y. Gong and X. Liu, "Generic text summarization using relevance measure and latent semantic analysis," in *Proc. of SIGIR*, pp. 19–25, 2001.

[10] J. Carbonell and J. Goldstein, "The use of MMR, diversity based reranking for reordering documents and producing summaries," in *Proc. of SIGIR*, pp. 335–336, 1998.

[11] X. Wan and J. Yang, "Multi-document summarization using cluster-based link analysis," in *Proc. of SIGIR*, pp. 299–306, 2008.

[12] G. Erkan and D. R. Radev, "LexRank: Graph-based lexical centrality as salience in text summarization", *Journal of Artificial Intelligent Research*, vol. 22, no. 1, pp. 457–479, 2004.

[13] H. Lin and J. Bilmes, "Multi-document summarization via budgeted maximization of submodular functions," in *Proc. of NAACL HLT,* pp. 912–920, 2010.

[14] K. Riedhammer *et al.*, "Long story short - Global unsupervised models for keyphrase based meeting summarization," *Speech Communication*, vol. 52, no. 10, pp. 801–815, 2010.

[15] K. Y. Chen *et al.*, "Extractive broadcast news summarization leveraging recurrent neural network language modeling techniques," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, no. 8, pp. 1322–1334, 2015.

[16] 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.

[17] S. H. Liu *et al.*, "Positional language modeling for extractive broadcast news speech summarization", in *Proc. of INTERSPEECH*, 2015.

[18] Y. Bengio *et al.*, "A neural probabilistic language model," *Journal of Machine Learning Research*, vol. 3, pp. 1137–1155, 2003.

[19] T. Mikolov *et al.*, "Efficient estimation of word representations in vector space," in *Proc. of ICLR*, pp. 1–12, 2013.

[20] G. Miller and W. Charles, "Contextual correlates of semantic similarity," *Language and Cognitive Processes*, vol. 6, no. 1, pp. 1–28, 1991.

[21] T. Mikolov *et al.*, "Distributed representations of words and phrases and their compositionality," in *Proc. of ICLR*, pp. 1–9, 2013.

[22] S. Brin and L. Page, "The anatomy of a large-scale hypertextual (web) search engine," in *Proceedings of the International Conference on the World Wide Web*, pp. 107–117, 1998.

[23] F. Morin and Y. Bengio, "Hierarchical probabilistic neural network language model," in *Proc. of AISTATS*, pp. 246–252, 2005.

[24] A. Mnih and K. Kavukcuoglu, "Learning word embeddings efficiently with noise-contrastive estimation," in *Proc. of NIPS*, pp. 2265–2273, 2013.

[25] M. J. Kusner *et al.*, "From word embeddings to document distances", in *Proc. of ICML*, vol. 37, 2015.

[26] Y. Rubner, C. Tomasi, and L. J. Guibas, "A metric for distributions with applications to image databases," in *Proc. of ICCV*, pp. 59–66, 1998.

[27] O. Pele, and M. Werman, "Fast and robust earth mover's distances," in *Proc. of ICCV*, pp. 460–467, 2009.

[28] X. Wan, "A novel document similarity measure based on earth mover's distance," *Information Sciences*, vol. 177, pp. 3718-3730, 2007.

[29] H. M. Wang *et al.*, "MATBN: A Mandarin Chinese broadcast news corpus," *International Journal of Computational Linguistics and Chinese Language Processing*, vol. 10, no. 2, pp. 219–236, 2005.

[30] C. Y. Lin, "ROUGE: Recall-oriented understudy for gisting evaluation." 2003 [Online]. Available: http://www.berouge.com/Pages/default.aspx.

[31] G. Heigold et al., "Discriminative training for automatic speech recognition: Modeling, criteria, optimization, implementation, and performance," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 58–69, 2012.

[32] P. B. Baxendale, "Machine-made index for technical literature-an experiment," *IBM Journal*, vol. 2, no. 4, pp. 354–361, 1958.

[33] M. Wasson, "Using leading text for news summaries: Evaluation results and implications for commercial summarization applications," in *Proc. of COLING*, pp. 1364–1368, 1998.