A Margin-based Discriminative Modeling Approach for Extractive Speech Summarization

Shih-Hung Liu[†*], Kuan-Yu Chen[*], Berlin Chen[#], Ea-Ee Jan[+]
Hsin-Min Wang[*], Hsu-Chun Yen[†], Wen-Lian Hsu[*]

[*] Institute of Information Science, Academia Sinica, Taiwan
E-mail: {journey, kychen, whm, hsu}@iis.sinica.edu.tw
[†] National Taiwan University, Taiwan
E-mail: yen@cc.ee.ntu.edu.tw
[#] National Taiwan Normal University, Taiwan
E-mail: berlin@ntnu.edu.tw
[+] IBM Thomas J. Watson Research Center, USA
E-mail: ejan@us.ibm.com

***Abstract*—The task of extractive speech summarization is to select a set of salient sentences from an original spoken document and concatenate them to form a summary, facilitating users to better browse through and understand the content of the document. In this paper we present an empirical study of leveraging various supervised discriminative methods for effectively ranking important sentences of a spoken document to be summarized. In addition, we propose a novel margin-based discriminative training (MBDT) algorithm that aims to penalize non-summary sentences in an inverse proportion to their summarization evaluation scores, leading to better discrimination from the desired summary sentences. By doing so, the summarization model can be trained with an objective function that is closely coupled with the ultimate evaluation metric of extractive speech summarization. Furthermore, sentences of spoken documents are embodied by a wide range of prosodic, lexical and relevance features, whose utilities are extensively compared and analyzed. Experiments conducted on a Mandarin broadcast news summarization task demonstrate the performance merits of our summarization method when compared to several well-studied state-of-the-art supervised and unsupervised methods.**

I. INTRODUCTION

With the prevalence of multimedia associated with spoken documents and the rapid advance in automatic speech recognition techniques [1, 2], research on extractive speech summarization has attracted an increasing interest in the speech processing community over the past decade [3-6]. Extractive speech summarization manages to select indicative sentences from an original spoken document according to a target summarization ratio and string them together to form a concise summary accordingly. By doing so, it can provide locations of salient speech segments alongside their corresponding transcripts for users to listen to and digest. In this paper, we focus exclusively on extractive speech summarization, even though we will typically omit the qualifier "extractive" hereafter.

Besides traditional unsupervised summarization methods [3-9], such as those based on document structural, linguistic or prosodic information, proximity or significance measures and relevance scores to identify salient sentences, machine-learning approaches with supervised training have drawn much attention and been applied with good success in a wide arrange of summarization tasks [3-9]. Specifically, the problem of speech summarization can be formulated as follows: Construct a ranking model (summarizer) that assigns a decision score (or a posterior probability) of being included in the summary to each sentence of a spoken document to be summarized. Then, important sentences are ranked and selected according to these scores [11-15]. As a simplest example, the summarizer can cast the speech summarization task as a two-class (summary/non-summary) sentence-classification problem: A spoken sentence with a set of indicative features is input to the summarizer and a decision score (or label) is then returned from it on the basis of these features [10].

Our work in this paper continues this general line of research and its main contributions are three-fold. First, we present an empirical study of leveraging various supervised discriminative methods for effectively ranking important sentences of a spoken document to be summarized. Second, we propose a novel margin-based discriminative training (MBDT) algorithm that aims to penalize non-summary sentences in an inverse proportion to their summarization evaluation scores, leading to better discrimination from the desired summary sentences. By doing so, the summarization model can be trained with an objective function that is closely linked to the ultimate evaluation metric of speech summarization. Finally, sentences of spoken documents are embodied by a wide range of prosodic, lexical and relevance features, whose utilities are extensively compared and analyzed.

The rest of this paper is organized as follows. Section II describes the theoretical underpinnings of various state-of-the-art supervised summarization methods we investigate and

compare in this paper. Section III elucidates the notion and the instantiation of the proposed margin-based discriminative training (MBDT) algorithm. After that, the experimental settings and a series of summarization experiments are presented in Sections IV and V. Finally, Section VI concludes this paper and suggests avenues for future work.

II. Supervised Summarization Methods

Without loss of generality, the sentence ranking strategy for extractive speech summarization can be stated as follows. Each sentence S_i in a spoken document to be summarized is associated with a set of M indicative features $\mathbf{X}_i = \{x_{i1}, \cdots, x_{im}, \cdots, x_{iM}\}$ (usually, represented in vector form) and a summarizer (or a real-valued ranking function) is employed to assign a decision (or importance) score to each sentence S_i according to its associated features $\mathbf{X}_i$. Sentences of the document in turn can be ranked and iteratively selected to be included into the summary based on their scores until the length limitation or a desired summarization ratio is reached. During the training phase, a set of training spoken documents $\mathbf{D} = \{D_1, \cdots, D_n, \cdots, D_N\}$, consisting of N documents and their corresponding handcrafted summary information, is used to train the supervised summarizer (or model).

In what follows, we describe four supervised summarizers that we will investigate for speech summarization, i.e., Support Vector Machines (SVM), Ranking SVM, Perceptron and the Global Conditional Log-linear model (GCLM). The former two have been well-studied in various text and speech summarization tasks, while for the latter two, to the best of our knowledge, there is still a dearth of work investigating them in the context of speech summarization.

A. Support Vector Machine (SVM)

An SVM summarizer is developed under the basic principle of structural risk minimization (SRM) in the statistical learning theory. If the dataset is linear separable, SVM attempts to find an optimal hyper-plane by utilizing a decision function that can correctly separate the positive and negative samples, and ensure the margin is maximal. In the nonlinear separable case, SVM uses kernel functions or defines slack variables to transform the problem into a linear discrimination problem. In this paper, we use the LIBSVM [1] toolkit to construct a binary SVM summarizer, and adopt the radial basis function (RBF) as the kernel function. The posterior probability of a sentence S_i being included in the summary class $\mathbf{S}$ can be approximated by the following sigmoid operation:

$$P(S_i \in \mathbf{S} \mid \mathbf{X}_i) \approx \frac{1}{1+\exp(\alpha \cdot g(S_i) + \beta)}, \quad (1)$$

where the weights α and β are optimized by the training set, and $g(S_i)$ is the decision score of the sentence S_i provided by the SVM summarizer. Once the SVM summarizer has been properly constructed, the sentences of a spoken document to be summarized can be ranked by their posterior probabilities of being in the summary class. The sentences with the highest probabilities are then selected and sequenced to form the final summary according to different summarization ratios.

Typically, the SVM summarizer is trained in the sense of reducing the classification errors of the summarizer made on the sentences of these training spoken document exemplars. It is anticipated that minimizing the classification errors caused by the summarizer would be equivalent to maximizing the lower bound of the summarization evaluation score (usually, the higher the score, the better the performance). In addition, the SVM summarizer in fact treats each training (summary or non-summary) sentence independently in estimating the corresponding model parameters. Theoretically, the training paradigm can be referred to as *point-wise* learning.

B. Ranking SVM

In contrast to SVM, Ranking SVM seeks to create a more rank- or preference-sensitive ranking function. It assumes there exists a set of ranks (or preferences) $L=\{l_1,l_2,\ldots,l_K\}$ in the output space, while in the context of speech summarization, the value of K, for example, can be simply set to 2 representing that a sentence can have the label of being either a summary (l_1) or a non-summary (l_2) sentence. The elements in the rank set have a total ordering relationship $l_1 \succ l_2 \succ \cdots \succ l_K$, where $\succ$ denotes a preference relationship. The training objective of Ranking SVM is to find a ranking function that can correctly determine the preference relation between any pair of sentences:

$$l(S_i) \succ l(S_j) \Leftrightarrow f(S_i) \succ f(S_j), \quad (2)$$

where $l(S_i)$ denotes the label of a sentence S_i and $f(S_i)$ denotes the decision value of S_i provided by Ranking SVM. As such, the corresponding training paradigm of Ranking SVM can be referred to as *pair-wise* learning. We refer to [16] for a more comprehensive and enjoyable discussion of Ranking SVM.

C. Perceptron

The Perceptron method that has been well-studied in natural language processing and speech recognition [17, 18] can also be adopted and formalized for speech summarization. The decision score that the Perceptron method gives to a candidate summary sentence S_i can be computed by

$$f(S_i) = \boldsymbol{\alpha} \cdot \mathbf{X}_i \quad (3)$$

where $\mathbf{X}_i$ is the feature vector used to characterize a candidate summary sentence S_i, and $\boldsymbol{\alpha}$ is the model parameter vector of the Perceptron method. Namely, in (3), a candidate summary sentence having a higher inner product value $\boldsymbol{\alpha} \cdot \mathbf{X}_i$ is more likely to be selected into the summary. The model parameter vector of Perceptron can be estimated by maximizing the accumulated squared score distances of all the training spoken documents defined as follows:

[1] http://www.csie.ntu.edu.tw/~cjlin/libsvm/index.html

$$F_{Perceptron}(\boldsymbol{\alpha}) = \frac{1}{2} \cdot \sum_{n=1}^{N} \sum_{S_R \in \text{Summ}_n} \left(f(S_R) - f(S_n^*) \right)^2, \quad (4)$$

where N is total training documents, $\mathbf{Summ}_n$ is the reference summary of the n-th training document D_n, S_R denotes a summary sentence in $\mathbf{Summ}_n$, and S_n^* is the non-summary sentence of D_n that has the highest decision score. After some algebraic manipulations, each component α_d of the parameter vector $\boldsymbol{\alpha}$ can be updated in an iterative manner using the following gradient descent formula:

$$\hat{\alpha}_d = \alpha_d + \eta \times \frac{\partial F_{Perceptron}(\boldsymbol{\alpha})}{\partial \alpha_d} \quad (5)$$

where η is a constant used to control the step size for parameter updating.

D. Global Conditional Log-linear Model (GCLM)

The GCLM method has its roots from speech recognition for re-ranking recognition hypotheses for better recognition accuracy [19, 20]. It also has the same sentence ranking function as the Perception method (*cf.* Eq. (3)), except that its model parameter vector $\boldsymbol{\alpha}$ is estimated by maximizing the following objective function:

$$F_{\text{GCLM}}(\boldsymbol{\alpha}) = \sum_{n=1}^{N} \sum_{S_R \in \text{Summ}_n} \log \frac{\exp(\boldsymbol{\alpha} \cdot \mathbf{X}_R)}{\sum_{S_j \in D_n} \exp(\boldsymbol{\alpha} \cdot \mathbf{X}_j)}, \quad (6)$$

By doing so, the GCLM method will maximize the posterior of the summary sentences (and thereby minimize the posterior of the non-summary sentences) of each given training spoken document.

III. Margin-based Discriminative Training

In this paper, we propose a novel margin-based discriminative training (MBDT) algorithm that additionally takes into account the ultimate evaluation metric of speech summarization when training a summarizer. In this way, the resulting summarizer can more effectively discriminate between summary sentences and non-summary sentences during the summarization process. To this end, the MBDT algorithm proceeds in two stages. In the first stage, MBDT conducts a training data selection procedure to select a subset of most confusing non-summary sentences S_i (i.e., to form a support set $\mathbf{Sup}_{S_R}$) for each summary sentence S_R in a training spoken document D_n that lies close to the decision boundary for speech summarization, which is expressed as follows:

$$\mathbf{Sup}_{S_R} = \{S_j \mid \tau_{S_R}(S_j) \le \varepsilon\}, \quad (7)$$

where ε is a tunable threshold, and $\tau_{S_R}(S_j)$ is the separation margin computed by

$$\tau_{S_R}(S_j) = f(S_R) - f(S_j). \quad (8)$$

In the second stage, MBDT attempts to define a training objective function that is closely coupled with the ultimate evaluation metric of speech summarization. Suppose that the model parameter vector of the summarizer is represented by $\boldsymbol{\alpha}$ and the summarizer also conducts ranking of important sentences using Eq. (3). Then, the model parameter vector $\boldsymbol{\alpha}$ of the summarizer can be estimated by maximizing the following objective function:

$$F_{\text{MBDT}}(\boldsymbol{\alpha}) = \frac{1}{2} \cdot \sum_{n=1}^{N} \sum_{S_R \in \mathbf{Summ}_n} \sum_{S_j \in \mathbf{Sup}_{S_R}} w(S_j) \left(f(S_R) - f(S_j) \right)^2, \quad (8)$$

where $w(S_j)$ is the weight of a (non-summary) sentence S_j which is defined as follows:

$$w(S_j) = 1 - Eval(S_j, \mathbf{Summ}_n) \quad (9)$$

where $Eval(S_j, \mathbf{Summ}_n)$ is a function that estimate the summarization performance of a sentence S_j of D_n by comparing S_j to the reference summary $\mathbf{Summ}_n$ of D_n with a desired evaluation metric, which will return a score ranging between 0 and 1 (again, the higher the value, the better the performance).

TABLE I
FEATURES USED IN THIS PAPER.

Prosodic Features	1. Pitch value (max, min, mean, diff) 2. Energy value (max, min, mean, diff)
Lexical Features	1. Number of named entities 2. Number of stop words 3. Bigram language model scores 4. Normalized bigram scores
Relevance Features	1. VSM 2. DLM 3. RM 4. SMM

IV. Experimental Setup

A. Features Characterizing Spoken Sentences

In this paper, we use a heterogeneous set of 16 indicative features to characterize a spoken sentence, including the lexical features, the prosodic features and the relevance features. Lexical features represent the linguistic characteristics. Prosodic features describe more about how things are said than what is said, and may provide additional important information for summarization. Lastly, relevance features evaluate the relevance between a spoken document and each one of its sentences. For each prosodic feature, the minimum, maximum, mean and difference values of a spoken sentence are extracted. In addition, the difference value is defined as the difference between the minimum and maximum values of the spoken sentence. Table I gives an outline of the different types of features used in this paper, where VSM (Vector Space Model) [21], DLM (Document Likelihood Measure) [22], RM (Relevance Model) [23, 24] and SMM (simple mixture model) [24, 25] are the relevance values output by the corresponding common unsupervised summarizers, and each is counted as a single summarization (relevance) feature respectively.

B. Speech and Language Corpora

The summarization dataset employed in this study is a broadcast news (MATBN) corpus collected by the Academia Sinica and the Public Television Service Foundation of Taiwan between November 2001 and April 2003 [26]. Each story contains the speech of one studio anchor, as well as several field reporters and interviewees. A subset of 205 broadcast news documents compiled between November 2001 and August 2002 was reserved for the summarization experiments. 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 proposed summarization method and the other state-of-the-art methods, while another subset of 100 documents selected from the rest is reserved as the training set.

Three subjects were asked to create summaries of the spoken documents as references (the gold standard) for evaluation. The reference summaries were generated by ranking the sentences in the manual transcript of each spoken document by importance without assigning a score to each sentence. For the assessment of summarization performance, we adopted the widely-used ROUGE metrics [27]. Three variants of the ROUGE metric were used to quantify the utility of the proposed methods. They are, respectively, the ROUGE-1 (unigram) metric, the ROUGE-2 (bigram) metric and the ROUGE-L (longest common subsequence) metric. All the experimental results reported hereafter are obtained by calculating the F-scores of these ROUGE metrics. 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.

V. Experiment Results

At the beginning, we assess the performance levels of the various supervised summarizers compared in this paper, i.e., SVM, Ranking SVM, Perceptron and GCLM. Notice here that all these summarizers are learned from the spoken documents of the training set along with their respective reference summaries, and then tested on the spoken documents of the evaluation set. The corresponding results of these four summarizers (in terms of ROUGE-1, ROUGE-2 and ROUGE-L metrics) are shown in Table II, where TD denotes the results obtained based on the manual transcripts of spoken documents and SD denotes the results using the speech recognition transcripts that may contain speech recognition errors. Furthermore, the results obtained by two other state-of-the-art unsupervised summarizers, i.e., the integer linear programming (ILP) method [28] and the submodularity-based method (Submodularity) [29] are also listed in Table II for reference.

Table II
Summarization results achieved by various summarization methods.

		ROUGE-1	ROUGE-2	ROUGE-L
TD	SVM	0.470	0.364	0.426
	Ranking SVM	0.490	0.391	0.447
	Perceptron	0.487	0.394	0.439
	GCLM	0.482	0.386	0.433
	MBDT	0.515	0.422	0.462
	ILP	0.442	0.337	0.401
	Submodularity	0.414	0.286	0.363
SD	SVM	0.383	0.245	0.342
	Ranking SVM	0.388	0.254	0.344
	Perceptron	0.393	0.259	0.352
	GCLM	0.380	0.250	0.342
	MBDT	0.393	0.264	0.353
	ILP	0.348	0.209	0.306
	Submodularity	0.332	0.204	0.303

Table III
Summarization results achieved by MBDT with different types of features to characterize spoken sentences.

		ROUGE-1	ROUGE-2	ROUGE-L
TD	MBDT+Pro	0.374	0.256	0.337
	MBDT+Lex	0.255	0.159	0.228
	MBDT+Rel	0.411	0.287	0.360
	MBDT+Pro+Lex	0.370	0.269	0.345
	MBDT+Lex+Rel	0.428	0.314	0.382
	MBDT+Pro+Rel	0.422	0.315	0.370
	MBDT+All	0.515	0.422	0.462
SD	MBDT+Pro	0.325	0.189	0.292
	MBDT+Lex	0.189	0.082	0.170
	MBDT+Rel	0.360	0.202	0.298
	MBDT+Pro+Lex	0.342	0.205	0.310
	MBDT+Lex+Rel	0.355	0.214	0.313
	MBDT+Pro+Rel	0.341	0.197	0.288
	MBDT+All	0.393	0.264	0.353

Several noteworthy observations can be drawn from Table II. First, for the TD case, Ranking SVM[2], Perceptron and GCLM tend to perform on par with one another when evaluated with the various ROUGE metrics. However, SVM yields inferior results as compared with the former three summarizers. This is mainly because that the training objective functions of the former three summarizers (i.e., Ranking SVM, Perceptron, GCLM) explicitly take into account the relatedness among summery and non-summary sentences. Therefore, they would have higher capability to distinguish between summery and non-summary sentences. Second, for the SD case, the performance gaps between SVM and the other three supervised summarizers are diminished, probably due to the fact that the dramatic performance degradation caused by imperfect speech recognition may overwhelm the relatively subtle performance differences among these supervised summarizers. Third, it comes as no

[2] http://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html

surprise that the two unsupervised summarizers (ILP and Submodularity) are apparently worse than the four supervised summarizers investigated in this paper.

In the second set of experiments, we evaluate the effectiveness of our proposed summarization method (i.e., MBDT), whose results are also shown in Table II. As can be seen, for the TD case, MBDT offers consistent and considerable improvements over all the aforementioned supervised and unsupervised methods in terms of the three ROGUE metrics. It reveals that additionally incorporating the knowledge about the ultimate evaluation metric of speech summarization into the training objective function of the summarizer can further boost its summarization performance. However, the improvements seem to be less pronounced for the SD case. Again, it is because that speech recognition errors will affect the faithful calculation of the estimated summarization performance of a given non-summary sentence (*cf.* Eq. (9)), resulting in an incorrect training objective function of MBDT.

In the third set of experiments, we analyze the contributions of the three types of features (i.e., prosodic features (denoted by Pro), lexical features (denoted by Lex) and relevance features (denoted by Rel)) and different combinations of them, which we use to characterize spoken sentences, on the final summarization performance, taking the MBDT summarizer as an example. The corresponding results are shown in Table IV. As expected, it is observed that using the relevance features (a kind of more elaborated features) in isolation can achieve the best performance than using the other two types of features. On the other hand, to our surprise, the prosodic features deliver summarization performance superior to the lexical features. One possible explanation is that the prosodic features indeed play a significant part in speech summarization, and they are less sensitive to the effect of imperfect speech recognition (for the SD case) compared to the lexical features. Furthermore, these three types of features are complementary to one another, since they can conspire to achieve the best performance (*cf.* MBDT+All in Table II) as compared to only using either one or two out of them for representing the spoken sentences.

VI. Conclusion and Outlook

In this paper, we have presented an empirical study to capitalize on several state-of-the-art supervised and unsupervised summarizers for speech summarization. In addition, we have proposed a novel margin-based discriminative training (MBDT) algorithm that has the ability to penalize non-summary sentences in an inverse proportion to their summarization evaluation scores during the model training phase, thereby resulting in a summarizer that can better discriminate the desired summary sentences from the non-summary ones. A series of experiments conducted on a broadcast news summarization task have demonstrated the performance merits of the MBDT-based summarizer when compared to several existing supervised summarizers. As to future work, we envisage to explore more sophisticated modeling techniques [30, 31] and training objective functions for speech summarization. We also plan to investigate more robust indexing mechanisms and confidence measures to alleviate the negative effect caused by imperfect speech recognition.

VII. Acknowledgement

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, NSC 101-2221-E-003-024-MY3, NSC 102-2221-E-003-014-, NSC 101-2511-S-003-057-MY3, NSC 101-2511-S-003-047-MY3 and NSC 103-2911-I-003-301.

References

[1] S. Furui, L. Deng, M. Gales, H. Ney. and K. Tokuda, "Fundamental technologies in modern speech recognition," IEEE Signal Processing Magazine, 29(6):16–17, 2012

[2] D. O'Shaughnessy, L. Deng and H. Li, "Speech information processing: Theory and applications," Proceedings of the IEEE, 101(5):1034–1037, 2013.

[3] S. Furui, T. Kikuchi, Y. Shinnaka and C. Hori, "Speech-to-text and speech-to-speech summarization of spontaneous speech," IEEE Transactions on Speech and Audio Processing, 12(4):401–408, 2004.

[4] K. McKeown, J. Hirschberg, M. Galley and S. Maskey, "From text to speech summarization," in *Proc. ICASSP*, pp. 997–1000, 2005.

[5] Y. Liu and D. Hakkani-Tur, "Speech summarization," in G. Tur and R. D. Mori [Ed], Spoken Language Understanding: Systems for Extracting Semantic Information from Speech, Wiley, 2011.

[6] A. Nenkova and K. McKeown, "Automatic summarization," Foundations and Trends in Information Retrieval, 5(2–3):103–233, 2011.

[7] H. Christensen, Y. Gotoh and S. Renals, "A cascaded broadcast news highlighter," IEEE Transactions on Audio, Speech and Language Processing, 16:151–161, 2008.

[8] J. Zhang and P. Fung, "Speech summarization without lexical features for Mandarin broadcast news," in *Proc. NAACL-HLT*, pp. 213–216, 2007.

[9] S. H. Lin, Y. T. Chen, H. M. Wang and B. Chen, "A comparative study of probabilistic ranking models for Chinese spoken document summarization," ACM Transactions on Asian Language Information Processing, 8(1): article 3, 2009.

[10] Y. T. Lo, S. H. Lin, B. Chen, "Constructing effective ranking models for speech summarization," in *Proc. ICASSP*, pp. 5053–5056, 2012.

[11] M. A. Fattah and F. Ren, "GA, MR, FFNN, PNN and GMM based models for automatic text summarization," Computer Speech & Language, 23(1):126–144, 2009.

[12] J. Kupiec, J. Pedersen and F. Chen, "A trainable document summarizer," in *Proc. SIGIR*, pp. 68–73, 1995.

[13] S. H. Lin, Y. M. Chang, J. W. Liu and B. Chen, "Leveraging evaluation metric-related training criteria for speech summarization," in *Proc. ICASSP*, pp. 5314–5317, 2010.

[14] B. Chen, S. H. Lin, Y. M. Chang and K. Y. Chen, "Extractive speech summarization using evaluation metric-related training criteria," Information Processing & Management, 49:1-12, 2013.

[15] M. Galley, "Skip-chain conditional random field for ranking meeting utterances by importance," in *Proc. EMNLP*, pp. 364–372, 2006.

[16] Y. Cao, J. Xu, T. Y. Liu, H. Li, Y. Huang and H. W. Hon, "Adapting ranking SVM to document retrieval," in *Proc. SIGIR*, pp. 186–193, 2006.

[17] M. Collins, "Discriminative training methods for hidden Markov models: theory and experiments with perceptron algorithms," in *Proc. EMNLP*, pp. 1-8, 2002

[18] Z. Zhou, J. Gao, F. K. Soong and H. Meng, "A comparative study of discriminative methods for reranking LVCSR N-best hypotheses in domain adaptation and generalization," in *Proc. ICASSP*, pp. 141-144, 2006.

[19] B. Roark, M. Saraclar and M. Collins, "Corrective language modeling for large vocabulary ASR with the perceptron algorithm," in *Proc. ICASSP*, pp. 749–752, 2004.

[20] B. Roark, M. Saraclar and M. Collins, "Discriminative n-gram language modeling," Computer Speech & Language, 21(2):373-392, 2007.

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

[22] Y. T. Chen, B. Chen and H. M. Wang, "A Probabilistic Generative Framework for Extractive Broadcast News Speech Summarization," IEEE Transactions on Audio, Speech and Language Processing, 17(1):95–106, 2009

[23] V. Lavrenko and W. B. Croft, "Relevance-based language models," in *Proc. SIGIR*, pp. 120–127, 2001.

[24] S. H. Liu, K. Y. Chen, Y. L. Hsieh, B. Chen, H. M. Wang, H. C. Yen and W. L. Hsu, "Effective Pseudo-relevance Feedback for Language Modeling in Extractive Speech Summarization," in *Proc. ICASSP*, pp. 3250–3254, 2014.

[25] C. X. Zhai and J. Lafferty, "Model-based feedback in the language modeling approach to information retrieval," in *Proc. CIKM*, pp. 403–410, 2001.

[26] H. M. Wang, B. Chen, J. W. Kuo and S. S. Cheng, "MATBN: A Mandarin Chinese broadcast news corpus," International Journal of Computational Linguistics and Chinese Language Processing, 10(2):219–236, 2005.

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

[28] K. Riedhammer, B. Favre, and D. Z. Hakkani-Tür,, "Long story short– Global unsupervised models for keyphrase based meeting summarization," Speech Communication, 52(10): 801–815, 2010.

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

[30] D. Li and D. Yu, *Deep Learning: Methods and Applications*, Foundations and Trends in Signal Processing, Now Publishers, June 2014.

[31] T. Mikolov, M. Karafiát, L. Burget, J. Cernocký and S. Khudanpur, "Recurrent neural network based language model," in *Proc. Interspeech*, pp. 1045–1048, 2010.