Supervised Spoken Document Summarization Based on Structured Support Vector Machine with Utterance Clusters as Hidden Variables

Sz-Rung Shiang [1], Hung-yi Lee [2], Lin-shan Lee [1]

[1] Graduate Institute of Electrical Engineering, National Taiwan University
[2] Research Center for Information Technology Innovation, Academia Sinica
r01921050@ntu.edu.tw, tlkagkb93901106@gmail.com, lslee@gate.sinica.edu.tw

10.21437/Interspeech.2013-626

Abstract

This paper presents a supervised approach for extractive summarization of spoken document considering utterance clusters in the documents as hidden variables. Utterances in important clusters may be jointly included in the summary, while those in less important clusters may be excluded as a whole. The summaries are therefore selected based on not only the conventional principle of including the most important utterances and minimizing the redundancy but also the hidden cluster structure in the document. The cluster structure of the documents is not known but can be inferred from the documents, and the summaries can be jointly obtained by the structured SVM learned from the training examples. Encouraging results were obtained on a lecture corpus in the preliminary experiments.

Index Terms: speech summarization, structured SVM, hidden variables.

1. Introduction

For extractive spoken document summarization, the summary is a subset of utterances automatically selected from the given spoken document [1]. A widely used unsupervised approach is the Maximum Marginal Relevance (MMR) method [2, 3], in which a greedy approach considering the balance between the importance of the selected utterances and the redundancy among the selected utterances is used for utterance selection. The unsupervised graph-based approach [4, 5, 6, 7] is also well-known and considers the overall relationship among the utterances of the spoken document using a graph.

Supervised learning [8, 9, 10, 11, 12, 13] has also been widely used, in which the task is very often treated as a binary classification problem regarding whether to include an utterance in the summary [14, 15]. With the availability of a set of documents and their reference summaries, the utterances in and not in the reference summaries are respectively taken as positive and negative examples in training the binary classifier, for example, a support vector machine (SVM) [16, 17, 18]. However, since most utterances in the document are not included in the summary, the imbalanced data may be a problem. Moreover, in this way the utterances are usually considered individually by the binary classifiers, although they are related in some way. To solve these problems, structured support vector machine (SVM) [19] was proposed to take the entire document as a whole during training process and directly generate an utterance subset for a given document as the summary instead of labeling positive or negative utterances [20].

Additionally, it has been found that document structure is very helpful in extractive spoken document summarization. ClusterRank [4] is a good example, in which the utterances in a document are first clustered, and then a graph-based approach is performed over the clusters. In this way, even though the importance of some utterances may not be directly identified based on their content; they can still be included in the summary if belonging to a meaningful cluster.

In this paper, we consider the structure or clusters of utterances of the spoken documents as hidden variables in structured SVM for supervised summarization. We assume utterances in important clusters tend to be jointly included in the summary while less important clusters of utterances may be excluded as a whole, so the utterances are selected not only based on the goal of including the most important utterances and minimizing the redundancy, but considering the utterance clusters as well. Because the utterance clusters are not directly observable in the spoken documents, they are jointly inferred with the summaries. A set of parameters regarding summary extraction, utterance clustering, and the relationship between the summaries and the clusters are all jointly learned from a set of training documents with reference summaries, although the clusters behind each training document are not available as hidden variables either. This is achieved by structured SVM with hidden variables [21, 22].

2. Proposed Approach

2.1. Previously proposed summarization based on Structured SVM

Based on the maximal marginal relevance (MMR) principle [2, 3], extractive spoken document summarization can be formulated as searching for the utterance subset s_d from a document d which maximizes the following objective function $F(d, s_d)$ [23, 24]:

$$\mathrm{F(d, s_d)} = \sum_{x_i \in s_d} R(x_i) - \lambda \sum_{x_i, x_j \in s_d} Sim(x_i, x_j)$$
$$s.t. \sum_{x_i \in s_d} L(x_i) \leq K \quad (1)$$

where $R(x_i)$ is the importance score of the utterance x_i in the subset s_d, $Sim(x_i, x_j)$ is the similarity between two utterances x_i and x_j, and $L(x_i)$ is the length of utterance x_i. So s_d is obtained by selecting the most important utterances while minimizing the redundancy under the constraint of a total length. It was shown previously that the parameter λ and weights parameters for evaluating $R(x_i)$ can be learned by the structured SVM from training data [20].

2.2. Summary Generation in Proposed Approach

In this paper, we introduce cluster structure into the structured SVM of (1) as shown in Fig.1, where x_i are utterances of a document d, h_k is an utterance cluster, H_d is the cluster set, and s_d is the summary. For simplicity, in Fig.1 and the experiments reported below, each utterance is clustered only with its neighbors as was done in ClusterRank

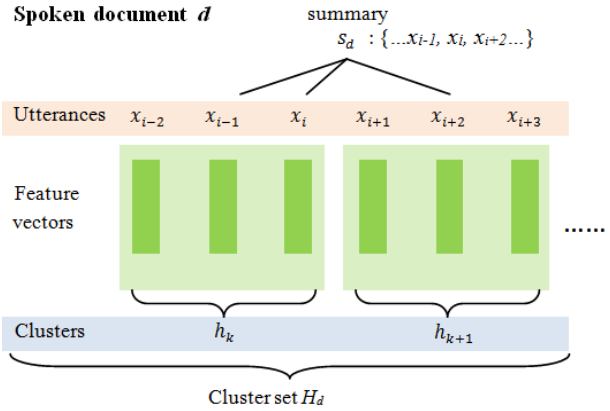

Figure 1: *Proposed approach: spoken document (d) and its utterances (x_i), clusters (h_k), cluster set (H_d) and summary (s_d).*

[4] (i.e. the utterance clustering here is simply the segmentation of the utterance sequence), although the proposed approach is not limited this way. Here the proposed approach searches within each document d for an utterance subset s_d (the summary) considering the cluster set H_d by maximizing the objective function $F(d, s_d, H_d)$ in (2).

$$F(d, s_d, H_d) = \sum_{x_i \in s_d} R(x_i) - \lambda \sum_{x_i, x_j \in s_d} Sim(x_i, x_j) + \sum_{h_k \in H_d} C(s_d, h_k) + \sum_{h_k \in H_d} S(h_k)$$

$$s.t. \sum_{x_i \in s_d} L(x_i) \le K \quad (2)$$

The first two terms in (2) are the same as in (1). Therefore the first two terms reward the utterance subset s_d which includes more important but not redundant utterances x_i. The third term in (2), $\sum_{h_k \in H_d} C(s_d, h_k)$, evaluates the utterance subset s_d as the summary considering the cluster set H_d with clusters h_k. For example, $C(s_d, h_k)$ will be large if all the utterances in the cluster h_k are included in or excluded from the summary s_d. $\sum_{h_k \in H_d} S(h_k)$ in the last term of (2) evaluates the quality of the cluster set H_d. For instance, $S(h_k)$ will be large if the utterances in h_k are highly similar to each other. Since the cluster set H_d is not observable for the document, the cluster set H_d and the summary s_d are inferred jointly during the summarization process. As a result, an utterance subset s_d with utterances x_i having high $R(x_i)$ and low redundancy $\sum_{x_i, x_j \in s_d} Sim(x_i, x_j)$ will not be taken as the summary, if there is no proper cluster set H_d (with high $\sum_{h_k \in H_d} S(h_k)$ in (2)) giving high $\sum_{h_k \in H_d} C(s_d, h_k)$ in (2).

More precisely, we can rewrite (2) and have (3) below:

$$F(d, s_d, H_d) = \sum_{x_i \in s_d} {\omega_0}^T \cdot F_0(x_i) - \lambda \sum_{x_i, x_j \in s_d} Sim(x_i, x_j) + \sum_{h_k \in H_d} {\omega_1}^T \cdot F_1(s_d, h_k) + \sum_{h_k \in H_d} {\omega_2}^T \cdot F_2(h_k) \quad (3)$$

where ω_0, ω_1, ω_2 are the weight vectors to be learned by the structured SVM, and $F_0(x_i)$, $F_1(s_d, h_k)$ and $F_2(h_k)$ are the respective feature vectors. $F_0(x_i)$ in $R(x_i)$ of the first term of (2)(3) is the feature vector indicating the importance of the utterance x_i, whose components may include such information of the utterance as the position, semantic topic probabilities, prosodic features and so on. Similarly, $F_1(s_d, h_k)$ in $C(s_d, h_k)$ of the third term of (2)(3) is the feature vector whose components indicate the relationship between the utterance subset s_d and the cluster h_k, and $F_2(h_k)$ in $S(h_k)$ of the last term of (2)(3) is the feature vector related to the quality of cluster h_k. For simplicity, we can rewrite (3) as (4):

$$F(d, s_d, H_d) = \omega^T \Phi(d, s_d, H_d) \quad (4)$$

where $\Phi(d, s_d, H_d)$ is the concatenation of the feature vectors $F_0(x_i)$, $F_1(s_d, h_k)$ and $F_2(h_k)$, and the scalar $Sim(x_i, x_j)$ in (3), and ω is the concatenation of ω_0, ω_1, $\omega_2$and -λ, which will be jointly learned as described in the next subsection.

2.3. Training Parameters in Proposed Approach

Let the training set be defined as, $T = \{(d_i, \hat{s}_{d_i}), i = 1,2,...n\}$,where $(d_i, \hat{s}_{d_i})$ is a training example with the reference summary $\hat{s}_{d_i}$ for the i-th document d_i. Ideally, if the oracle (best) cluster set $\hat{H}_{d_i}$ for the training document d_i was known, the goal of training would be to find the weight vector ω such that the summary $\hat{s}_{d_i}$ and cluster set $\hat{H}_{d_i}$ maximize $\omega^T \Phi(d_i, s_{d_i}, H_{d_i})$ in (4). In fact, although the cluster set behind each document is unknown, the reference summary s_{d_i} reveals how the utterances should be clustered because the summaries should have relationships with the clusters as defined in (4). Therefore, it is possible to infer the best cluster set H_{d_i} for the document d_i from the given reference summary $\hat{s}_{d_i}$. Practically, the training process can search over all possible cluster sets H_{d_i} of d_i and find the best H_{d_i} which maximizes $\omega^T \Phi(d_i, s_{d_i}, H_{d_i})$ when s_{d_i} is $\hat{s}_{d_i}$ (that is, to find the best cluster set H_{d_i} given the reference summary $\hat{s}_{d_i}$), and then take this best H_{d_i} as the oracle cluster set $\hat{H}_{d_i}$.

With the oracle cluster set $\hat{H}_{d_i}$for each training document, the following optimization function learns the parameters ω in (4) using structured SVM as listed below in (5) with constraints (6) and (7):

$$min_{\omega, \xi} \frac{1}{2} ||\omega||^2 + C \sum_i \xi_i \quad (5)$$

$$\forall i, \quad \hat{H}_{d_i} = \underset{H_i}{argmax}\, \omega^T \cdot \Phi(d_i, \hat{s}_{d_i}, H_{d_i}) \quad (6)$$

$$\forall i, \forall s_{d_i} \neq \hat{s}_{d_i}, \forall H_i, \quad \omega^T \cdot \Phi(d_i, \hat{s}_{d_i}, \hat{H}_{d_i}) - \omega^T \cdot \Phi(d_i, s_{d_i}, H_{d_i}) \ge \Delta(s_{d_i}, \hat{s}_{d_i}) - \xi_i, \quad \xi_i \ge 0 \quad (7)$$

where C is the trade-off parameter similar to that in SVM, and the norm term is the parameters to be learned. In the constraint of (6), $\hat{H}_{d_i}$ is the utterance cluster set (or hidden variable) of d_i maximizing $\Phi(d_i, \hat{s}_{d_i}, H_{d_i})$ as mentioned above. In the constraint (7), for each training document d_i, it is required that $\omega^T \cdot \Phi(d_i, \hat{s}_{d_i}, \hat{H}_{d_i})$ exceeds $\omega^T \cdot \Phi(d_i, s_{d_i}, H_{d_i})$ for any other subset s_{d_i} and cluster set H_{d_i} by at least a document-dependent margin $\Delta(s_{d_i}, \hat{s}_{d_i})$ which is a function of s_{d_i} and $\hat{s}_{d_i}$. In the experiments below, $\Delta(s_{d_i}, \hat{s}_{d_i})$ is defined as $1 - ROUGE(s_{d_i})$, where $ROUGE(s_{d_i})$ is the ROUGE 1-F measure [25] if s_{d_i} is considered as summary and $\hat{s}_{d_i}$ is the reference summary.

Other choices of the loss function are certainly possible. This margin becomes larger when s_{d_i} is far from the reference summary $\hat{s}_{d_i}$. Hence, when s_{d_i} is a poorer summary, the margin will be larger, or it would be less probable to be mistaken as the summary. This document-dependent margin is padded with a document-dependent slack variable ξ_i whose sum over the training set is minimized in (5). The optimization problem in (5) is complicated with a huge number of constraints, but it can be solved using the Concave-Convex Procedure [22, 26].

3. Features parameters used in the experiments

3.1. $F_0(x_i)$: Features for an utterance

3.1.1. Semantic features

We used Probability Latent Semantic Analysis (PLSA) [27] to analyze the semantics of each utterance. PLSA uses a set of latent topic variables $\{T_k, k = 1,2,\ldots,K\}$ to characterize the "term-utterance" co-occurrence relationships. PLSA training over all the utterances in the spoken document archive yields $\{P(w_i \mid T_k),\ i = 1, 2, \ldots, V,\ k = 1, 2, \ldots, K\}$, the probability of observing a term w_i in an utterance given the topic T_k, and $\{P(T_k \mid x),\ k = 1, 2, \ldots, K\}$, the mixture weight of topic T_k given the transcription of an utterance *x*. The latter *K* probabilities $\{P(T_k \mid x),\ k = 1, 2, \ldots, K\}$ are then taken as the *K* components of $F_0(x_i)$. $K = 32$ in the experiments reported below.

3.1.2. Similarity to the whole document

The similarity measure $Sim(x, d)$ between an utterance x and the whole document d is defined as the average of the similarity measure $Sim(x, x')$ between x and each utterance x' in d. $Sim(x, x')$ is the cosine similarity between the vector representations v and v' for x and x', where the vector representations can be either lexical-based or PLSA-based. For lexical-based similarity, each component of *v* corresponds to a word in the lexicon, whose value is the term frequency (tf) weighted by the inverse document frequency (idf) for the term. For PLSA-based similarity, the dimension of v is the number of latent topics *K* as mentioned above in Subsection 3.1.1, and the value of each component in *v* is simply $P(T_k|x)$, $k = 1, 2, \ldots, K$ from PLSA.

3.1.3. Prosodic feature

It is well known that prosody is very helpful to spoken document summarization [15, 28, 29]. We used 60 prosodic features for each utterance, 27 related to pause and syllable duration, 13 to energy and 20 to pitch. The details are left out for space limitation.

3.1.4. Key term related feature

Key terms appearing in an utterance certainly indicate the importance of the utterance. The key terms which are automatically extracted by the approach developed previously for course lectures [30] were used here. There were 2 feature parameters defined for an utterance based on key terms:

- The number of key terms in an utterance.
- Assume a key term occurring the first time in a document gave more new information than the same key term appearing latter on. Hence, in an utterance the number of key terms occurring the first time in the document was taken as a feature parameters.

3.1.5. Other feature parameters

- Utterance length: number of words and/or Chinese characters in the utterance's transcription.
- Normalized utterance position: i/N for the i-th utterance in a document with N utterances.
- Significance scores: The sum of the significance scores $I(w)$ for all terms w in the utterance's transcription. $I(w) = tf(w) \times idf(w)$, where $tf(w)$ is the frequency count of w in the whole document, and $idf(w)$ is the inverse document frequency of w [31]; or $I'(w) = tf(w)/LTE(w)$, where $LTE(w)$ is the latent topic entropy for *w* based on PLSA [32].

3.2. $F_1(s_d, h_k)$: Features for relationship between the cluster and the summary

- Inclusion completeness: There are two components related to this property. One is 1.0 if all the utterances of the cluster h_k are included in the summary s_d, whereas the other is 1.0 if all the utterances in the h_k cluster are not included in the summary s_d.
- Consecutiveness: For each utterance x_i in the cluster h_k, $n(x_i, s_d, h_k)$ is 1 if one or more of its neighbor utterances within h_k are also included in the summary s_d, and 0 otherwise. The parameter is then the sum of $n(x_i, s_d, h_k)$ over all utterances x_i in h_k which are included in s_d.

3.3. $F_2(h_k)$: Features for the quality of the cluster h_k

- Average similarities for all pairs of utterances within the cluster h_k, and PLSA-based cosine similarity as mentioned in section 3.1.2 was used here.
- Similarity between the cluster h_k and the document *d*: average of the PLSA-based cosine similarity $Sim(x, d)$ for all the utterances *x* in the cluster h_k.

4. Experiments

4.1. Corpus and experiment setup

The corpus used in this research was the lectures of coding-switching nature for courses offered at National Taiwan University with a total length of 45 hours produced by a single instructor. The lectures were given in the host language of Mandarin Chinese but with many technical terms uttered in the guest language of English embedded in the Mandarin utterances. The corpus was segmented into 193 documents based on the slides used, and the average document length was about 17.5 minutes. We divided the corpus into two parts. 12 hours of speech was used for acoustic model training, and we also used its manual transcriptions for language model adaptation. The remaining 33 hours were used for testing. One-best ASR transcriptions were used for summarization, with ASR accuracy (for Chinese characters / English words) being 88.0%.

Only 40 documents with reference summaries in the testing set of the corpus were used for the experiments below. The reference summaries were utterances selected from the documents generated by graduate students who had taken

Table 1. *F-measures of ROUGE-1, 2, and L for short(10%) and long(30%) summaries using different approaches.*

		unsupervised		Supervised			
constraint	Evaluation Measure	(a) trivial baseline	(b) MMR	(c) individual-utterance SVM	(d) structured SVM	(e) proposed (without inclusion completeness)	(f) proposed (All)
10%	ROUGE-1	0.3839	0.3966	0.4117	0.4315	0.4363	0.4406
	ROUGE-2	0.1438	0.1777	0.1761	0.2162	0.2329	0.2208
	ROUGE-L	0.3732	0.3983	0.4057	0.4229	0.4285	0.4333
30%	ROUGE-1	0.5134	0.5484	0.5372	0.5624	0.5628	0.5657
	ROUGE-2	0.3114	0.3380	0.3354	0.3500	0.3688	0.3627
	ROUGE-L	0.5102	0.5445	0.5335	0.5577	0.5591	0.5616

these courses. Each spoken document has 3 short (10% summarization ratio) and 3 long (30%) reference summaries. Summarization ratio of 10% means the length (number of Chinese characters plus English words in the manual transcriptions) of the summaries didn't exceed 10% of the whole spoken documents. The 40 documents were divided into 4 folds. For each trial, 3 folds (30 documents) were used as training data and the remaining 1 fold (10 documents) for testing. When training a summarizer generating short summaries (10% ratio), the short reference summaries were used, while the long reference summaries were used for long summaries. A training document with 3 reference summaries was regarded as 3 training examples. ROUGE F-measures [25] were used to evaluate the summarization results.

4.2. Baseline approaches

Both unsupervised and supervised approaches were taken as baseline for comparison. A trivial approach of selecting the longest utterances under the summarization ratio was taken as the first unsupervised baseline. The second unsupervised baseline was the well-known MMR [2] method with the lexical similarity between an utterance x_i and the whole document regarded as the importance score. The first supervised baseline classified each utterance individually as positive (in summary) or negative (not in summary) using standard SVM [33]. The second supervised baseline used the structured SVM proposed previously [20] as in (1) of section 2.1, or only the first two terms of (2) in section 2.2 were considered.

4.3. Experimental result

Table 1 lists the F-measures of ROUGE-1, 2 and L with summarization ratio of 10% (upper part) and 30% (lower part) using different approaches. Columns (a) and (b) are for trivial and MMR unsupervised baselines, while columns (c), (d), (e) and (f) are for supervised baselines. Columns (c) and (d) are respectively for SVM with individual utterances and structured SVM proposed previously, and columns (e) and (f) are our proposed method. Column (e) is the results utilizing all the features in Section 3 but excluding the "inclusion completeness" mentioned in section 3.2, while all features were used in column (f). In the experiments in columns (e) and (f), the number of utterances in a cluster was limited to between 3 and 10. We found that MMR remarkably outperformed the trivial baselines (columns (b) vs (a)), but individual-utterance SVM did not surpass MMR even though it was supervised due to the problem described in Section 1(columns (c) vs (b)). On the other hand, structured SVM outperformed the other baselines owing to the solution to the considerations for the relationships between utterances and the balance between the importance and redundancy which was jointly learned(columns (d) vs (a), (b), (c)). Our proposed approach made progress compared to structured SVM by considering the cluster structure (columns (e), (f) vs (d)). This shows that the document structure is informative and useful. When comparing the results without/with the consideration of inclusion completeness (columns (e) vs (f)), we could find that inclusion completeness improves ROUGE 1-F and L-F scores. Since ROUGE 1-F was taken as loss function in constraint (7), the system thus learned to prevent from generating summaries with poor ROUGE 1-F (ROUGE L-F was also improved because ROUGE L-F was highly correlated to ROUGE 1-F), while ROUGE 2-F was not considered during learning.

For the analysis of weight vector ω learned in (5), we noted that weight for consecutiveness was negative in 10% summarization ratio while positive in 30% summarization ratio. This reflected that long summaries consisted of some consecutive utterances and short summaries did not feature the characteristic because of the limited length constraint. The weights for inclusion completeness were all positive in 10% and 30% summaries which meant that the utterances in a cluster are included or excluded together in the summary. For the weights related to the quality of clusters in Section 3.3, the weight for similarity within a cluster was a large positive number in 10% and 30% summarization compared to the other weights, obviously because it had to do with the quality of the cluster. On the other hand, the weight for similarity between cluster and document was close to zero, probably because irrelevant clusters were also helpful to summarization if they could be completely excluded. That is, the utterances in a cluster should be highly similar to each other, but the similarity between the cluster and the document is not really related to the cluster quality.

5. Conclusions

In this paper, we proposed a supervised spoken document summarization approach, which could learn the cluster structure of the document, the importance of the utterances, and the balance for the redundancy between utterances at the same time. The proposed approach based on the clusters was shown to offer improvements over the previously proposed approach with structured SVM.

6. References

[1] Yang Liu and Dilek Hakkani-Tur, "Spoken Language Understanding Systems for Extracting Semantic Information from speech", chapter 13, pp. 357 – 396, Wiley, 2011.

[2] Shasha Xie and Yang Liu, "Using corpus and knowledge-based similarity measure in maximum marginal relevance for meeting summarization," in ICASSP, 2008.

[3] Jaime Carbonell and Jade Goldstein, "The use of mmr, diversity-based reranking for reordering documents and producing summaries," in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in*

[4] Nikhil Garg, Benoit Favre, Korbinian Reidhammer, Dilek Hakkani-T¨ur, "ClusterRank: A Graph Based Method for Meeting Summarization", in *Interspeech*, 2009.

[5] Hui Lin, J. Bilmes, and Shasha Xie, "Graph-based submodular selection for extractive summarization," in *ASRU*, 2009.

[6] Yun-Nung Chen, Yu Huang, Ching-Feng Yeh, and Lin-Shan Lee, "Spoken lecture summarization by random walk over a graph constructed with automatically extracted key terms," in *Interspeech*, 2011.

[7] Yun-Nung Chen and Florian Metze, "Integrating intra-speaker topic modeling and temporal-based inter-speaker topic modeling in randomwalk for improved multi-party meeting summarization," in *Interspeech*, 2012.

[8] Anne Hendrik Buist, Wessel Kraaij, and Stephan Raaijmakers, "Automatic summarization of meeting data: A feasibility study," in *Proc. Meeting of Computational Linguistics in the Netherlands (CLIN)*, 2004.

[9] Jian Zhang and Pascale Fung, "Speech summarization without lexical features for mandarin broadcast news," in *Proceedings of the Human Language Technology Conference of the NAACL*, 2007, pp. 213–216.

[10] Shih-Hsiang Lin, Berlin Chen, and Hsin-Min Wang, "A comparative study of probabilistic ranking models for chinese spoken document summarization," *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 8, pp. 3:1–3:23, 2009.

[11] J.J. Zhang, R.H.Y. Chan, and P. Fung, "Extractive speech summarization by active learning," in *ASRU*, 2009.

[12] Michel Galley, "A skip-chain conditional random field for ranking meeting utterances by importance," in *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, 2006.

[13] Shasha Xie and Yang Liu, "Improving supervised learning for meeting summarization using sampling and regression," *Computer Speech & Language*, vol. 24, pp. 495 – 514, 2010.

[14] J. Zhang, H. Y. Chan, P. Fung, and L. Cao, "A comparative study on speech summarization of broadcast news and lecture speech," in *Interspeech*, 2007.

[15] S. Xie, D. Hakkani-Tur, B. Favre, and Y. Liu, "Integrating prosodic features in extractive meeting summarization," in *ASRU*, 2009.

[16] Vladimir N. Vapnik, "The Nature of Statistical Learning Theory", Springer, 1995.

[17] Thorsten Joachims, "Making Large-Scale SVM Learning Practical", LS8-Report, 24, Universität Dortmund, LS VIII-Report, 1998.

[18] Thorsten Joachims, "Learning to Classify Text Using Support Vector Machines", Dissertation, Kluwer, 2002.

[19] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun, "Support vector machine learning for interdependent and structured output spaces," in *ICML*, 2004.

[20] Hung-yi Lee, Yu-yu Chou, Yow-Bang Wang, Lin-shan Lee, "Supervised Spoken Document Summarization Jointly Considering Utterance Importance and Redundancy by Structured Support Vector Machine", in *Interspeech*, 2012.

[21] Yang Wang, and Greg Mori, "Max-Margin Hidden Conditional Random Fields for Human Action Recognition", in *CVPR*, 2009.

[22] C.-N. Yu and T. Joachims, "Learning Structural SVMs with Latent Variables", in *ICML*, 2009.

[23] Ryan McDonald, "A study of global inference algorithms in multi-document summarization," in Proceedings of the 29th European conference on IR research, 2007.

[24] D. Gillick, K. Riedhammer, B. Favre, and D. Hakkani-Tur, "A global optimization framework for meeting summarization," in *ICASSP*, 2009.

[25] C. yew Lin, "Rouge: A package for automatic evaluation of summaries," in *Workshop on Text Summarization Branches Out*, 2004.

[26] Yuille, A., & Rangarajan, A, "The Concave-Convex Procedure", *Neural Computation*, 15, 915, 2003.

[27] Thomas Hofmann, "Probabilistic latent semantic analysis," in *UAI*, 1999.

[28] Sameer Maskey and Julia Hirschberg, "Comparing lexical, acoustic/prosodic, structural and discourse features for speech summarization," in *Interspeech*, 2005.

[29] Sameer Maskey and Julia Hirschberg, "Summarizing speech without text using hidden markov models," in *Proceedings of the Human Language Technology Conference of the NAACL*, 2006.

[30] Yun-Nung Chen, Yu Huang, Hung-Yi Lee, and Lin-Shan Lee, "Unsupervised two-stage keyword extraction from spoken documents by topic coherence and support vector machine," in *ICASSP*, 2012

[31] Sadaoki Furui, Tomonori Kikuchi, Yousuke Shinnaka, and Chiori Hori,"Speech-to-text and speech-to-speech summarization of spontaneous speech," *IEEE Trans. on Speech and Audio Processing*, vol. 12, no. 4, pp. 401–408, 2004.

[32] Sheng-Yi Kong and Lin-Shan Lee, "Semantic analysis and organization of spoken documents based on parameters derived from latent topics," *Audio, Speech, and Language Processing, IEEE Transactions* on, vol. 19, pp. 1875-1889, 2011.

[33] T. Joachims, "Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning", B. Schölkopf and C. Burges and A. Smola (ed.), *MIT-Press*, 1999.