# Active Learning of Extractive Reference Summaries for Lecture Speech Summarization

**Justin Jian Zhang and Pascale Fung**
Human Language Technology Center
Department of Electronic and Computer Engineering
University of Science and Technology (HKUST)
Clear Water Bay,Hong Kong
{zjustin,pascale}@ece.ust.hk

## Abstract

We propose using active learning for tagging extractive reference summary of lecture speech. The training process of feature-based summarization model usually requires a large amount of training data with high-quality reference summaries. Human production of such summaries is tedious, and since inter-labeler agreement is low, very unreliable. Active learning helps assuage this problem by automatically selecting a small amount of unlabeled documents for humans to hand correct. Our method chooses the unlabeled documents according to the similarity score between the document and the comparable resource—PowerPoint slides. After manual correction, the selected documents are returned to the training pool. Summarization results show an increasing learning curve of ROUGE-L F-measure, from 0.44 to 0.514, consistently higher than that of using randomly chosen training samples.

**Index Terms**: active learning, summarization

## 1 Introduction

The need for the summarization of classroom lectures, conference speeches, political speeches is ever increasing with the advent of remote learning, distributed collaboration and electronic archiving. These user needs cannot be sufficiently met by short abstracts. In recent years, virtually all summarization systems are extractive - compiling bullet points from the document using some saliency criteria. Reference summaries are often manually compiled by one or multiple human annotators (Fujii et al., 2008; Nenkova et al., 2007). Unlike for speech recognition where the reference sentence is clear and unambiguous, and unlike for machine translation where there are guidelines for manual translating reference sentences, there is no clear guideline for compiling a good reference summary. As a result, one of the most important challenges in speech summarization remains the difficulty to compile, evaluate and thus to learn what a good summary is. Human judges tend to agree on obviously good and very bad summaries but cannot agree on borderline cases. Consequently, annotator agreement is low. Reference summary generation is a tedious and low efficiency task. On the other hand, supervised learning of extractive summarization requires a large amount of training data of reference summaries. To reduce the amount of human annotation effort and improve annotator agreement on the reference summaries, we propose that active learning (selective sampling) is one possible solution.

Active learning has been applied to NLP tasks such as spoken language understanding (Tur et al., 2005), information extraction (Shen et al., 2004), and text classification (Lewis and Catlett, 1994; McCallum and Nigam, 1998; Tong and Koller, 2002). Different from supervised learning which needs the entire corpus with manual labeling result, active learning selects the most useful examples for labeling and requires manual labeling of training dataset to re-train model.

In this paper, we suggest a framework of reference summary annotation with relatively high inter labeler agreement based on the rhetorical structure in presentation slides. Based on this framework, we further propose a certainty-based active learning method to alleviate the burden of human annotation of training data.

The rest of this paper is organized as follows: Section 2 depicts the corpus for our experiments, the extractive summarizer, and outlines the acoustic/prosodic, and linguistic feature sets for representing each sentence. Section 3 depicts how to



*Proceedings of the 2nd Workshop on Building and Using Comparable Corpora, ACL-IJCNLP 2009*, pages 23–26,
Suntec, Singapore, 6 August 2009. ©2009 ACL and AFNLP

compile reference summaries with high inter labeler agreement by using the RDTW algorithm and our active learning algorithm for tagging extractive reference summary. We describe our experiments and evaluate the results in Section 4. Our conclusion follows in Section 5.

## 2 Experimental Setup

### 2.1 The Corpus

Our lecture speech corpus (Zhang et al., 2008) contains 111 presentations recorded from the NCMMSC2005 and NCMMSC2007 conferences for evaluating our approach. The manual transcriptions and the comparable corpus—PowerPoint slides are also collected. Each presentation lasts for 15 minutes on average. We select 71 of the 111 presentations with well organized PowerPoint slides that always have clear sketches and evidently aligned with the transcriptions. We use about 90% of the lecture corpus from the 65 presentations as original unlabeled data $U$ and the remaining 6 presentations as held-out test set. We randomly select 5 presentations from $U$ as our seed presentations. Reference summaries of the seed presentations and the presentations of test set are generated from the PowerPoint slides and presentation transcriptions using RDTW followed by manual correction, as described in Section 3.

### 2.2 SVM Classifier and the Feature Set

While (Ribeiro and de Matos, 2007) has shown that MMR (maximum marginal relevance) approach is superior to feature-based classification for summarizing Portuguese broadcast news data, another work on Japanese lecture speech drew the opposite conclusion (Fujii et al., 2008) that feature-based classification method is better. Therefore we continue to use the feature-based method in our work. We consider the extractive summarization as a binary classification problem, we predict whether each sentence of the lecture transcription should be in a summary or not. We use Radial Basis Function (RBF) kernel for constructing SVM classifier, which is provided by LIBSVM, a library for support vector machines (Chang and Lin, 2001). We represent each sentence by a feature vector which consists of acoustic features: duration of the sentence, average syllable Duration, F0 information features, energy information features; and linguistic features: length of the sentence counted by word and TFIDF information features, as shown in (Zhang et al., 2008). We then build the SVM classifier as our summarizer based on these sentence feature vectors.

## 3 Active Learning for Tagging Reference Summary and Summarization

Similar to (Hayama et al., 2005; Kan, 2007), we have previously proposed how presentation slides are used to compile reference summaries automatically (Zhang et al., 2008). The motivations behind this procedure are:

- presentation slides are compiled by the authors themselves and therefore provide a good standard summary of their work;
- presentation slides contain the hierarchical rhetorical structure of lecture speech as the titles, subtitles, page breaks, bullet points provide an enriched set of discourse information that are otherwise not apparent in the spoken lecture transcriptions.

We propose a Relaxed Dynamic Time Warping (RDTW) procedure, which is identical to Dynamic Programming and Edit Distance, to align sentences from the slides to those in the lecture speech transcriptions, resulting in automatically extracted reference summaries.

We calculate the similarity scores matrix $Sim = (s_{ij})$, where $s_{ij} = similarity(Sent_{trans}[i], Sent_{slides}[j])$, between the sentences in the transcription and the sentences in the slides. We then obtain the distance matrix $Dist = (d_{ij})$, where $d_{ij} = 1 - s_{ij}$. We calculate the initial warp path P: $P = (p_1^{ini}, ..., p_n^{ini}, ..., p_N^{ini})$ by DTW, where $p_n^{ini}$ is represented by sentence pair$(i_n^{ini}, j_n^{ini})$: one from transcription, the other from slides. Considering that the lecturer often doesn't follow the flow of his/her slides strictly, we adopt Relaxed Dynamic Time Warping (RDTW) for finding the optimal warp path, by the following equation.

$$\begin{cases} i_n^{opt} = i_n^{ini} \\ j_n^{opt} = \underset{j=j_n^{ini}-C}{\overset{j_n^{ini}+C}{\operatorname{argmin}}} d_{i_n^{opt},j} \end{cases} \quad (1)$$

We consider the transcription sentences on this path as reference summary sentences. We then obtain the optimal path $(p_1^{opt}, ..., p_n^{opt}, ..., p_N^{opt})$, where $p_n^{opt}$ is represented by $(i_n^{opt}, j_n^{opt})$ and $C$

is the capacity to relax the path. We then select the sentences $i_n^{opt}$ of the transcription whose similarity scores of sentence pairs: $(i_n^{opt}, j_n^{opt})$, are higher than the pre-defined threshold as the reference summary sentences. The advantage of using these summaries as references is that it circumvents the disagreement between multiple human annotators.

We have compared these reference summaries to human-labeled summaries. When asked to "select the most salient sentences for a summary", we found that inter-annotator agreement ranges from 30% to 50% only. Sometimes even a single person might choose different sentences at different times (Nenkova et al., 2007). However, when instructed to follow the structure and points in the presentation slides, inter-annotator agreement increased to 80%. The agreement between automatically extracted reference summary and humans also reaches 75%. Based on this high degree of agreement, we generate reference summaries by asking a human to manually correct those extracted by the RDTW algorithm. Our reference summaries therefore make for more reliable training and test data.

For a transcribed presentation $D$ with a sequence of recognized sentences $\{s_1, s_2, ..., s_N\}$, we want to find the sentences to be classified as summary sentences by using the salient sentence classification function $c()$. In a probabilistic framework, the extractive summarization task is equivalent to estimating $P(c(\overrightarrow{s}_n) = 1|D)$ of each sentence $s_n$, where $\overrightarrow{s}_n$ is the feature vector with acoustic and linguistic features of the sentence $s_n$.

We propose an active learning approach where a small set of transcriptions as seeds with reference summaries, created by the RDTW algorithm and human correction, are used to train the seed model for the summarization classifier, and then the classifier is used to label data from a unlabel pool. At each iteration, human annotators choose the unlabeled documents whose similarity scores between the extracted summary sentences and the PowerPoint slides sentences are top-N highest for labeling summary sentences. Formally, this approach is described in Algorithm 1.

Given document $D$: $\{s_1, s_2, ..., s_N\}$, we calculate the similarity score between the extracted summary sentences: $\{s_1^{'}, s_2^{'}, ..., s_K^{'}\}$ and the PowerPoint slide sentences: $\{ppts_1, ppts_2, ..., ppts_L\}$, by equation 2.

$$Score_{sim}(D) = \frac{1}{K}\sum_{n=1}^{K}\sum_{j=1}^{L} Sim(s_n^{'}, ppts_j) \quad (2)$$

## 4 Experimental Results and Evaluation

**Algorithm 1** Active learning for tagging extractive reference summary and summarization

**Initialization**

For an unlabeled data set: $U_{all}$, $i = 0$
**(1)** Randomly choose a small set of data $X\{i\}$ from $U_{all}$; $U\{i\} = U_{all} - X\{i\}$
**(2)** Manually label each sentence in $X\{i\}$ as summary or non-summary by RDTW and human correction and save these sentences and their labels in $L\{i\}$

**Active Learning Process**

**(3)** $X\{i\} = null$
**(4)** Train the classifier $M\{i\}$ using $L\{i\}$
**(5)** Test $U\{i\}$ by $M\{i\}$
**(6)** Calculate similarity score of given document $D$ between the extracted summary sentences and the PowerPoint slides sentences by equation 2
**(7)** Select the documents with top-five highest similarity scores from $U\{i\}$
**(8)** Save selected samples into $X\{i\}$
**(9)** Manually correct each sentence label in $X\{i\}$ as summary or non-summary
**(10)** $L\{i+1\} = L\{i\} + X\{i\}$
**(11)** $U\{i+1\} = U\{i\} - X\{i\}$
**(12)** Evaluate $M\{i\}$ on the testing set $E$
**(13)** $i = i + 1$, and repeat from **(3)** until $U\{i\}$ is empty or $M\{i\}$ obtains satisfying performance
**(14)** $M\{i\}$ is produced and the process ends

We start our experiments by randomly choosing six documents for manual labeling. We gradually increase the training data pool by choosing five more documents each time for manual correction. We carry out two sets of experiments for comparing our algorithm and random selection. We evaluate the summarizer by ROUGE-L (summary-level Longest Common Subsequence) F-measure (Lin, 2004).

The performance of our algorithm is illustrated by the increasing ROUGE-L F-measure curve in Figure 1. It is shown to be consistently higher than

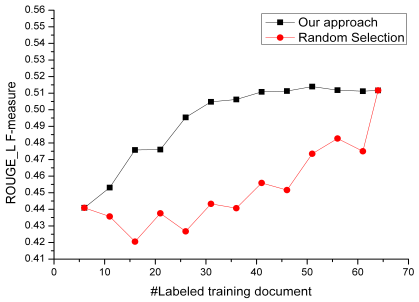


Figure 1: Active learning vs. random selection

using randomly chosen samples. We also find that by using only 51 documents for training, the performance of the summarization model achieved by our approach is better than that of the model trained by *random selection* using all 65 presentations (0.514 vs. 0.512 ROUGE-L F-measure). This shows that our active learning approach requires 22% less training data. Besides, acoustic features can improve the performance of active learning of speech summarization. Without acoustic features, our summarizer only performs 0.47 ROUGE-L F-measure.

## 5 Conclusion and Discussion

In this paper, we propose using active learning reduce the need for human annotation for tagging extractive reference summary of lecture speech summarization. We use RDTW to extract sentences from transcriptions according to PowerPoint slides, and these sentences are then hand corrected as reference summaries. The unlabeled documents are selected whose similarity scores between the extracted summary sentences and the PowerPoint slides sentences are top-N highest for labeling summary sentences. We then use an SVM classifier to extract summary sentences. Summarization results show an increasing learning curve of F-measure, from 0.44 to 0.514, consistently higher than that of using randomly chosen training data samples. Besides, acoustic features play a significant role in active learning of speech summarization. In our future work, we will try to apply different criteria, such as uncertainty-based or committee-based criteria, for selecting samples to be labeled, and compare the effectiveness of them.

## 6 Acknowledgements

This work is partially supported by GRF612806 of the Hong Kong RGC.

## References

C.C. Chang and C.J. Lin. 2001. LIBSVM: a library for support vector machines. *Software available at http://www.csie. ntu. edu. tw/cjlin/libsvm*, 80:604–611.

Y. Fujii, K. Yamamoto, N. Kitaoka, and S. Nakagawa. 2008. Class Lecture Summarization Taking into Account Consecutiveness of Important Sentences. In *Proceedings of Interspeech*, pages 2438–2441.

T. Hayama, H. Nanba, and S. Kunifuji. 2005. Alignment between a technical paper and presentation sheets using a hidden markov model. In *Active Media Technology, 2005.(AMT 2005). Proceedings of the 2005 International Conference on*, pages 102–106.

M.Y. Kan. 2007. SlideSeer: A digital library of aligned document and presentation pairs. In *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, pages 81–90. ACM New York, NY, USA.

D.D. Lewis and J. Catlett. 1994. Heterogeneous uncertainty sampling for supervised learning. In *Proceedings of the Eleventh International Conference on Machine Learning*, pages 148–156. Morgan Kaufmann.

C.Y. Lin. 2004. Rouge: A Package for Automatic Evaluation of Summaries. *Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)*, pages 25–26.

A. McCallum and K. Nigam. 1998. Employing EM in Pool-based Active Learning for Text Classification. In *Proceedings of ICML*, pages 350–358.

A. Nenkova, R. Passonneau, and K. McKeown. 2007. The Pyramid Method: Incorporating human content selection variation in summarization evaluation. *ACM Transactions on Speech and Language Processing (TSLP)*, 4(2).

R. Ribeiro and D.M. de Matos. 2007. Extractive Summarization of Broadcast News: Comparing Strategies for European Portuguese. *Lecture Notes in Computer Science*, 4629:115.

D. Shen, J. Zhang, J. Su, G. Zhou, and C.L. Tan. 2004. Multi-criteria-based Active Learning for Named Entity Recognition. In *Proceedings of 42th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics Morristown, NJ, USA.

S. Tong and D. Koller. 2002. Support vector machine active learning with applications to text classification. *The Journal of Machine Learning Research*, 2:45–66.

G. Tur, D. Hakkani-Tr, and R. E. Schapiro. 2005. Combining Active and Semi-supervised Learning for Spoken Language Understanding. *Speech Communications*, 45:171–186.

J.J. Zhang, S. Huang, and P. Fung. 2008. RSHMM++ for extractive lecture speech summarization. In *IEEE Spoken Language Technology Workshop, 2008. SLT 2008*, pages 161–164.