Alignment of Spoken Utterances with Slide Content for Easier Learning with Recorded Lectures using Structured Support Vector Machine (SVM)

Han Lu[1], *Sheng-syun Shen*[1], *Sz-Rung Shiang*[2], *Hung-yi Lee*[3], *Lin-shan Lee*[1,2,3]

[1] Electrical Engineering, National Taiwan University
[2] Graduate Institute of Electrical Engineering, National Taiwan University
[3] Graduate Institute of Communication Engineering, National Taiwan University

b99901108@ntu.edu.tw, b99901107@ntu.edu.tw, b97901031@ntu.edu.tw,
tlkagkb93901106@gmail.com, lslee@gate.sinica.edu.tw

Abstract

This paper reports the first known effort to automatically align the spoken utterances in recorded lectures with the content of the slides used. Such technologies will be very useful in Massive Open On-line Courses (MOOCs) and various recorded lectures as well as many other applications. We propose a set of approaches considering the problem that words helpful for such alignment are sparse and noisy, and the assumption that the presentation of a slide is usually smooth and top-down across the slide. This includes utterance clustering, entropy-based word filtering, reliability-propagated word-based matching, and the structured support vector machine (SVM) learning from local and global features. Initial experimental results with the lectures in a course offered in National Taiwan University showed very encouraging results as compared to the baseline approaches.

Index Terms: alignment, structured SVM, global features

1. Introduction

With the fast increasing Massive Open On-line Courses (MOOCs) and the widely available recorded conference lectures, it is possible today for people worldwide to learn desired knowledge from recorded courses and lectures via Internet, which is creating a globalized learning-on-demand environment. In many of such recorded lectures, very often it is not recorded which part of the slide each spoken utterance is referring to, although this can be done if the lecturer really intends to do so. But such alignment is very helpful to learners when listening to the recorded lectures.

This paper reports the first known effort to handle the above problem, i.e. to automatically align the spoken utterances with the presentation slide content as shown in Figure 1. In Figure 1, the content of a slide $\boldsymbol{S}$ is divided into sections $\boldsymbol{P} = \{p_i | i = 1,2,\dots,|\boldsymbol{P}|\}$ primarily based on the subtitles on the slide, with the corresponding utterance set $\boldsymbol{U}$ for the slide, $\boldsymbol{U} = \{u_i | i = 1,2,\dots,|\boldsymbol{U}|\}$. The goal here is to align each u_i in $\boldsymbol{U}$ with a section p_j in $\boldsymbol{S}$ which u_i is referring to, or to obtain a set of aligned pairs,

$$A(\boldsymbol{U},\boldsymbol{S}) = \{< u_i, p_{b_i} > | i = 1,2,\dots,|\boldsymbol{U}|\}, \quad (1)$$

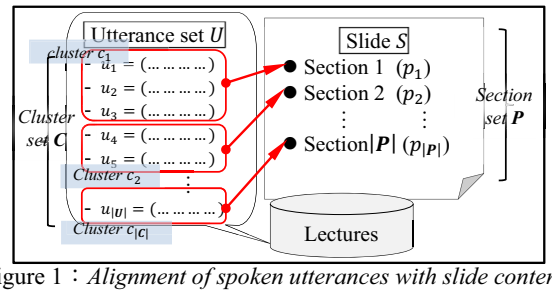

Figure 1：*Alignment of spoken utterances with slide content*

where b_i is the index j for the section p_j which u_i is aligned to. Here $\boldsymbol{U}$ is first divided into consecutive utterance clusters $\boldsymbol{C} = \{c_j | j = 1,2,\dots,|\boldsymbol{C}|\}$, each including one to several utterances. For example, in Figure 1 $c_1 = \{u_1, u_2, u_3\}$ is the first cluster including the first three utterances, etc. By properly aligning each cluster c_j of utterances with a section of the slide, the learner can feel much easier in learning with the recorded lectures. Efficient solution to this problem is useful not only to MOOCs and recorded lectures, but in many other applications such as spoken content retrieval and question answering based on spoken content.

Previously reported works related to slide alignment were primarily focused on aligning presentation slides (or plus the spoken part) with the sections of the corresponding text document such as technical papers or teaching materials. In other words, the unit of alignment in these cases is a slide, and the alignment is over a whole presentation, such as aligning the individual slides in a talk with the paper for the talk. But the work reported here is for the alignment "within a slide", or a cluster of several utterances aligned with a section of the slide. Most of the previously reported works tackled the alignment problem by matching the words [1, 2]. Some work further considered jointly text and images [3], while other works investigated the similarity measures, query expansion [4], and the short segment matching [5]. Still some other related works included information extraction from slides [6] and alignment between corpora [7].

The alignment within a slide between utterances clusters and sections of the slide is in general difficult, because a slide may be presented by a large number of utterances, while each utterance includes only limited number of words. So an utterance may include a keyword of the slide which appears repeatedly in various parts of the slide, while many other words in the utterances very probably never appear in the whole slide. This is referred to as the sparse/noisy word problem here, and is why approaches directly matching the words are not very useful. So we propose to divide the utterances into consecutive clusters $\{c_j | j = 1,2,\dots,|\boldsymbol{C}|\}$ and the slide content into sections $\{p_i | i = 1,2,\dots,|\boldsymbol{P}|\}$ and try to align c_j with p_i. The words in a cluster can be less sparse than an utterance, but more noisy. The approaches here are also benefited by the sequential smoothness assumption for presentation of slides, i.e., the presenter usually tends to finish a section of the slide before moving to another section, and very often follows the top-down section sequence in the slide, although not always.

The proposed approach is shown in Figure 2. The pre-processing includes utterance clustering and entropy-based word filtering. The output can then be used for both the reliability-propagated word-based matching and the structured SVM [8, 9, 10, 11, 12, 13]. The structured SVM can be trained either with a set of human-labeled data, or the output of the

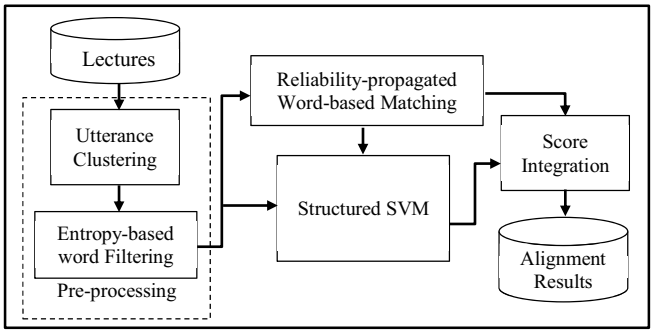

Figure 2： *The proposed approach*

reliability-propagated word-based matching. In the latter case, the whole process can be completely unsupervised. The scores of the reliability-propagated word-based matching and the structured SVM are then integrated.

2. Proposed Approach

We consider the problem defined in Figure 1. Here, in the initial work we assume each utterance u_i is the one-best recognition output or a word sequence, although probably better represented as a lattice.

2.1. Pre-Processing

This is to try to deal with the sparse/noisy word problem.

2.1.1. Utterance Clustering

By dividing all utterances into consecutive clusters as shown in Figure 1, a cluster can have less sparse words (although possibly more noisy), while aligning the utterances clusters with the sections of the slides may meet the sequential smoothness assumption to a certain degree. We simply calculate at every utterance boundary the cosine similarity between the two utterances on both sides based on tf-idf [14, 15, 16] vectors for words, with utterances taken as the documents for idf evaluation. Those boundaries with the similarity below a threshold TH_{sim} determined with a development set is then taken as the cluster boundary. This produces the cluster set $\boldsymbol{C}$ as in Figure 1,

$$\boldsymbol{C} = \{c_i | i = 1,2, \dots |\boldsymbol{C}|\}, \tag{2}$$

where c_i is the i-th cluster. In this way the desired alignment $A(\boldsymbol{U}, \boldsymbol{S})$ in (1) is reduced to

$$A(\boldsymbol{C}, \boldsymbol{S}) = \{< c_i, p_{a_i} > | i = 1,2, \dots, |\boldsymbol{C}|\}, \tag{3}$$

where a_i is the index j for section p_j which c_i is aligned to.

2.1.2. Entropy-based word Filtering

Here we remove some noisy words from the utterances. The basic idea is that if a word appears uniformly in all clusters, it is not useful in alignment even if it is a keyword. If a word appears frequently only in one or few clusters but not in other clusters, it will be useful in alignment. This concept is exactly the entropy useful in many areas [17, 18]. Let $f(w, i)$ be the term frequency of a word w in a cluster c_i, and $P(w, i) = f(w, i) / \sum_{k=1}^{|\boldsymbol{C}|} f(w, k)$ is the percentage of the word w appearing in the cluster c_i out of the entire slide $\boldsymbol{S}$. The entropy of the word w over the cluster set $\boldsymbol{C}$ is then

$$E(w|\boldsymbol{C}) = -\sum_{i=1}^{|\boldsymbol{C}|} p(w, i)\, logP(w, i). \tag{4}$$

Higher entropy $E(w|\boldsymbol{C})$ indicates the word w is less useful in alignment. So words with $E(w|\boldsymbol{C})$ above a threshold TH_{ent} determined with a development set is therefore simply deleted.

2.2. Reliability-propagated Word-based Matching

Given each utterance cluster c_i with noisy words deleted we first evaluate the term frequency vector (idf not used here since the words already selected by entropy), and then compute the cosine similarity with the tf-idf vectors for all sections (sections taken as documents in idf evaluation) of the slide, giving $\{Sim(c_i, p_j), j = 1,2, \dots, |\boldsymbol{P}|\}$. Naturally c_i can be aligned to the section p_j maximizing $Sim(c_i, p_j)$, i.e.,

$$a_i = arg\, max_j\, Sim(c_i, p_j), \tag{5}$$

where a_i is the index for the section which c_i is aligned to as in (3). However, the sparse/noisy word problem makes $Sim(c_i, p_j)$ in (5) less reliable, so we enhance it as below.

We define a cluster c_i to be reliable if the ratio of the maximum and second maximum of $Sim(c_i, p_j)$ exceeds a threshold TH_{rel} determined with a development set, or

$$\frac{max_j\, Sim(c_i, p_j)}{2nd\, max_j\, Sim(c_i, p_j)} \geq TH_{rel}, \tag{6}$$

in which cases $Sim(c_i, p_j)$ can be used directly in (5). Otherwise $Sim(c_i, p_j)$ should be enhanced in the following way.

When a cluster c_i is not reliable because (6) is not satisfied, let c_{i-k} represents the nearest reliable cluster before c_i satisfying (6) which is aligned to a section $p_{a_{i-k}}$ based on (5), and c_{i+l} represents the nearest reliable cluster after c_i aligned to $p_{a_{i+l}}$. It is then more likely that the unreliable cluster c_i is referring to a section close to $p_{a_{i-k}}$ or $p_{a_{i+l}}$ based on the sequential smoothness assumption. This implies $Sim(c_i, p_{a_{i-k}})$ and $Sim(c_i, p_{a_{i+1}})$ for this unreliable cluster c_i may be relatively more reliable, and the farther a section p_j is from $p_{a_{i-k}}$ or $p_{a_{i+1}}$, the less reliable $Sim(c_i, p_j)$ is. This leads to the reliability propagation weighting scheme in the following.

$$Sim'(c_i, p_j) = \frac{1}{2}[r_{i-k}(j) + r_{i+l}(j)] Sim(c_i, p_j), \tag{7}$$

$$r_{i-k}(j) = \begin{cases} 1, & \text{if j} = \text{a}_{\text{i-k}} \\ 1 - \text{b}, & \text{if j} = \text{a}_{\text{i-k}} \pm 1 \\ 1 - \text{b} - \text{md}, & \text{if j} = \text{a}_{\text{i-k}} \pm \text{m}, \text{m} \neq 1, \end{cases} \tag{8}$$

where $r_{i-k}(j)$ is the reliability propagation weight based on the nearest reliable cluster c_{i-k} before c_i, b, d are two discounting factors, $0 < b, d < 1.0$, and m is an integer. $r_{i+l}(j)$ is similarly defined as in (8). The values of $r_{i-k}(j)$ in (8) imply $Sim(c_i, p_j)$ is assumed reliable for $p_j = p_{a_{i-k}}$, but slightly discounted if p_j is next to $p_{a_{i-k}}$, and further discounted if p_j is farther away from $p_{a_{i-k}}$. The weight in (7) implies the value of $Sim(c_i, p_j)$ is enhanced by the two reliable clusters on both sides with reliability propagated from $p_{a_{i-k}}$ and $p_{a_{i+1}}$. $Sim'(c_i, p_j)$ in (7) is then the enhanced similarity to be used in (5) for those clusters not satisfying (6). The parameters b, d in (8) can be tuned with a development set.

2.3. Structured Support Vector Machine (SVM)

Here, we consider the alignment as an optimization problem, in which the desired alignment $A(\boldsymbol{C}, \boldsymbol{S})$ in (3) is the one maximizing the following objective function $F[A(\boldsymbol{C}, \boldsymbol{S})]$ among all possible alignments,

$$F[A(\boldsymbol{C}, \boldsymbol{S})] = \sum_{c_i \in \boldsymbol{C}} L(c_i, p_{a_i}) + G[A(\boldsymbol{C}, \boldsymbol{S})], \tag{9}$$

where $L(c_i, p_{a_i})$ is the local objective function for aligning c_i with p_{a_i}, and $G[A(\boldsymbol{C},\boldsymbol{S})]$ is the global objective function considering the whole alignment structure. We express $L(c_i, p_{a_i})$ and $G[A(\boldsymbol{C},\boldsymbol{S})]$ as linear functions with weight vectors $\overrightarrow{w_l}$ and $\overrightarrow{w_g}$ to be learned based on some training set:

$$L(c_i, p_{a_i}) = \overrightarrow{w_l} \cdot f_l(c_i, p_{a_i}), \quad (10)$$

$$G[A(\boldsymbol{C},\boldsymbol{S})] = \overrightarrow{w_g} \cdot f_g[A(\boldsymbol{C},\boldsymbol{S})], \quad (11)$$

where $f_l(c_i, p_{a_i})$ is the local feature vector representing the alignment relationship between c_i and p_{a_i}, and $f_g[A(\boldsymbol{C},\boldsymbol{S})]$ is the global feature vector representing the whole alignment structure. These features will be explained in detail in section 2.4 below. The local objective function encourages better individual alignment between c_i and p_{a_i}, while the global objective function encourages better overall alignment considering all alignment pairs $< c_i, p_{a_i} >, i = 1,2,\dots,|\boldsymbol{C}|$ in $A(\boldsymbol{C},\boldsymbol{S})$ jointly.

The above optimization problem can be solved using structured SVM, assuming the availability of a training set, $\{[< \boldsymbol{C}_i, \boldsymbol{S}_i >, \hat{A}(\boldsymbol{C}_i, \boldsymbol{S}_i)] | i = 1,2,\dots,n\}$, where $< \boldsymbol{C}_i, \boldsymbol{S}_i >$ is the i-th training example for slide $\boldsymbol{S}_i$ and the corresponding utterance cluster set $\boldsymbol{C}_i$, and $\hat{A}(\boldsymbol{C}_i, \boldsymbol{S}_i)$ is the reference alignment relationship for $< \boldsymbol{C}_i, \boldsymbol{S}_i >$. As will be shown below, the training set can be either a manually labeled set or those obtained with the pre-processing and reliability-propagated word-based matching as described above. In the latter case the structured SVM can be actually trained in a completely unsupervised way. With the training set, the goal here is to jointly learn the weight vectors $\overrightarrow{w_l}$ and $\overrightarrow{w_g}$ such that for every training example $< \boldsymbol{C}_i, \boldsymbol{S}_i >$ the reference alignment $\hat{A}(\boldsymbol{C}_i, \boldsymbol{S}_i)$ gives the highest objective function score $F[A(\boldsymbol{C}_i, \boldsymbol{S}_i)]$ among all possible alignments $A(\boldsymbol{C}_i, \boldsymbol{S}_i)$ using structured SVM [9, 10, 11, 12, 13]:

$$\min_{\overrightarrow{w_l}, \overrightarrow{w_g}} \frac{1}{2}\left(\|\overrightarrow{w_l}\|^2 + \|\overrightarrow{w_g}\|^2\right) + \frac{\alpha}{n}\sum_{i=1}^{n} \epsilon_i, \quad (12)$$

$$s.t. \forall i, \forall A(\boldsymbol{C}_i, \boldsymbol{S}_i), A(\boldsymbol{C}_i, \boldsymbol{S}_i) \neq \hat{A}(\boldsymbol{C}_i, \boldsymbol{S}_i):$$

$$F[\hat{A}(\boldsymbol{C}_i, \boldsymbol{S}_i)] - F[A(\boldsymbol{C}_i, \boldsymbol{S}_i)] \geq l[A(\boldsymbol{C}_i, \boldsymbol{S}_i)] - \epsilon_i, \epsilon_i \geq 0.$$

The constraints in (12) require that for each training example $< \boldsymbol{C}_i, \boldsymbol{S}_i >$, the differences between the objective function scores of the reference alignment $\hat{A}(\boldsymbol{C}_i, \boldsymbol{S}_i)$ and any other possible alignment $A(\boldsymbol{C}_i, \boldsymbol{S}_i)$ are larger than a margin $l[A(\boldsymbol{C}_i, \boldsymbol{S}_i)]$ padded by a per-instance slack of ϵ_i, where $l[A(\boldsymbol{C}_i, \boldsymbol{S}_i)]$ is a loss function when $A(\boldsymbol{C}_i, \boldsymbol{S}_i)$ is mistaken as the alignment. Hence, when $A(\boldsymbol{C}_i, \boldsymbol{S}_i)$ is a poorer alignment, the margin will be larger, or there would be less chance for it to be mistaken as the alignment. In the experiments reported below we define $l(A(\boldsymbol{C}_i, \boldsymbol{S}_i))$ as $1 - acc(A(\boldsymbol{C}_i, \boldsymbol{S}_i))$, where $acc(A(\boldsymbol{C}_i, \boldsymbol{S}_i))$ is the alignment accuracy or recall, i.e., the percentage of the alignment pairs $< c_k, p_j >$ in the reference alignment $\hat{A}(\boldsymbol{C}_i, \boldsymbol{S}_i)$ which also appear in $A(\boldsymbol{C}_i, \boldsymbol{S}_i)$. α in (12) is a constant for tradeoff between the slack variable ϵ_i and the norm of the weighting vectors to be learned. The optimization problem of (12) is a quadratic programming problem with huge number of constraints, but an approximate solution can be found in reasonable time with the cutting plane algorithm by selecting a set of active constraints out of all constraints [12].

2.4. Features used in the structured SVM

Here we describe the local feature vector $f_l(< c_i, p_j >)$ in (10) and the global feature vector $f_g[A(\boldsymbol{C},\boldsymbol{S})]$ in (11).

2.4.1. Local Features

The local features in the local feature vector $f_l(< c_i, p_j >)$ include the following:

- Cosine similarity between c_i and p_j based on the tf-idf.
- Number of distinct words that co-occur in c_i and p_j.
- $Sim(c_i, p_j) \cdot \delta(j, a_{i-k})$ and $Sim(c_i, p_j) \cdot \delta(j, a_{i+l})$. Assuming c_i is not reliable as defined in (6), and c_{i-k} and c_{i+l} represent the nearest reliable clusters before and after c_i as explained below (6), so a_{i-k} and a_{i+l} are the indices for the sections c_{i-k} and c_{i+l} are aligned to. $\delta(m,n) = 1$ if m=n and 0 else.
- $Sim(c_{i-1}, p_j) \cdot Sim(c_{i-1}, c_i)$ and $Sim(c_{i+1}, p_j) \cdot Sim(c_{i+1}, c_i)$, where $Sim(c_{i-1}, c_i)$ and $Sim(c_{i+1}, c_i)$ are the cosine similarity between c_i and its neighboring clusters on both sides.
- Number of distinct words that co-occur in c_i and p_j but not occurring in any other sections.

2.4.2. Global Features

The global features in the global feature vector $f_g[A(\boldsymbol{C},\boldsymbol{S})]$ include the following:

- Number of crossed alignment: number of alignment pairs $< c_i, p_{a_i} >$ and $< c_j, p_{a_j} >$ such that $i < j$ but $a_i > a_j$. This is the number of times that the presentation order is reversed or the sequential smoothness assumption is violated.
- Number of pairs $< c_i, p_{a_i} >$ in $A(\boldsymbol{C},\boldsymbol{S})$ such that $p_{a_i} \neq p_{a_{i+1}}$ normalized with $|\boldsymbol{P}|$. This is the number of times c_{i+1} is aligned to a section different from the one c_i is aligned to. For a slide with $|\boldsymbol{P}| = 5$, the most desirable sequential smoothness gives 80% for this feature, which means 4 times of section switching on the slide top-down. This feature penalizes departures from such a situation.
- Normalized squared length difference: $\sum_{i=1\sim|\boldsymbol{P}|}\{\text{length}(p_i) - \bar{n}[p_i, A(\boldsymbol{U},\boldsymbol{S})]\}^2$, where $length(p_i)$ is the total number of words in the section p_i normalized to the total number of words in the slide, and $\bar{n}[p_i, A(\boldsymbol{U},\boldsymbol{S})]$ is the number of utterances aligned to p_i in $A(\boldsymbol{U},\boldsymbol{S})$ normalized to the total number of utterances in $\boldsymbol{U}$. This feature assumes a longer section (with more words) should be explained with more utterances, therefore ideally this parameter should be close to zero. A larger value implies worse alignment quality.

2.5. Score Integration

The scores obtained with reliability-propagated word-based matching (section 2.2) and structured SVM (section 2.3, 2.4) can be integrated as in (13),

$$I[A(\boldsymbol{C},\boldsymbol{S})] = F[A(\boldsymbol{C},\boldsymbol{S})] + \lambda \sum_{i=1}^{|\boldsymbol{C}|} Sim(c_i, p_{a_i}), \quad (13)$$

where $Sim(c_i, p_{a_i})$ is the enhanced similarity in (7) if c_i is not reliable, or simply the similarity if c_i is reliable, and λ is a constant determined over a development set.

3. Experiments

3.1. Corpus and Experimental Setup

We used the lectures for a course offered at National Taiwan University with a total length of 45 hours as the corpus for this research, along with a total of 193 slides completely in English. The spoken lectures were, however, in Mandarin-English code-switching style, i.e. , the utterances were primarily in the host language of Mandarin, while the special terms and some other popularly used English words were produced in the guest language of English embedded in the Mandarin utterances [19].

We deleted all Chinese words in the one-best recognition output and considered only English words, except the utterance clustering used both Chinese and English words. When an utterance cluster is completely empty because it includes only Chinese words, we used the alignment for the nearest cluster which is not empty as its alignment.

The 45 hours of spoken lectures were segmented into 193 spoken documents based on the 193 slides. 12 hours of them were used for acoustic model training and in-domain language model adaptation. The accuracy for the one-best ASR transcriptions for the remaining 33 hours was 88.0% [20]. Only 38 spoken documents for 38 slides out of this testing set were used for alignment experiment below, for which relevance alignments were generated manually for each utterance.

We used 4 documents out of the 38 as the development set for determining all the parameters needed. The rest 34 documents were divided into 4 sets (9, 9, 8 and 8 documents) and used in 4-fold cross validation for testing the alignment accuracy, i.e. , testing each set using the structured SVM trained with the other three sets, and so on. For supervised approach we used the reference alignments of the training set to train the structured SVM. For completely unsupervised approach, we used the alignment results generated by the reliability-propagated word-based matching to train the structured SVM. In both cases the utterance clustering and entropy-based word filtering were performed on the utterances first, and the structured SVM then tried to learn the alignment between the automatically generated clusters and the slide sections.

Two baseline approaches were compared to here. Baseline 1 aligned each utterance to a section randomly. Baseline 2 performed the alignment for each utterance simply based on the similarity evaluated with tf-idf without utterance clustering, entropy-based word filtering, or any other approaches.

3.2. Experimental Results

All the results reported below are averaged alignment accuracy for utterances. It turned out that in average there were only 3.204 sections per slide for the corpus tested here, which gave 34.01% average accuracy for Baseline 1 of random alignment. The results are listed in Table 1. Column (a) is the results of Baseline 2 using cosine similarity with tf-idf but not any other processes, which is actually significantly better than Baseline 1 of random alignment (34.01%), indicating that tf-idf is really useful although relatively simple.

Columns (b), (c) and (d) are for completely unsupervised approaches, respectively for the reliability-propagated word-based matching, structured SVM (trained with output of (b)), and the score integration. The results of supervised approaches trained with human-labeled reference alignment are listed in columns (e) and (f), respectively using structured SVM alone (column (e)) or with score integration (column (f)).

We noted that the reliability-propagated word-based matching was in fact much better than Baseline 2 (columns (b) vs (a)), so the approaches presented in section 2.1 and 2.2 was really helpful. The unsupervised structured SVM was only slightly better (columns (c) vs (b)), probably because it was trained using the result of (b), therefore couldn't do too much differently from what it learned from. The score integration was actually much better than its individual components (columns (d) vs (b), (c)), apparently because the two component approaches were quite different and complementary to each other, although one of them learned the output of the other.

The supervised training of the structured SVM was able to offer much better results than unsupervised approaches (columns (e) vs (b) (c)), and the score integration was even better (columns (f) vs (e)), obviously because the structured SVM can not only learn from the individual alignments, but jointly consider the global features and local features.

On the other hand, the results with ASR output (2nd row) were actually not too far from those using manual transcriptions (1st row) with almost the same trend, probably because the ASR accuracy for the single-speaker lectures was not very low. This indicated that the approaches proposed here could be useful for real lectures (with alignment accuracy close to or exceeding 70% for all approaches in columns (b)-(f)), if the recognition accuracy was not very low. Furthermore, we note that the score integration results for unsupervised approaches (column (d)) were actually very close to (for manual transcriptions in 1st row) or even better than (for ASR output in 2nd row) the supervised structured SVM (column (e)). This not only verified the score integration was very strong in combining the strength of two very different component approaches, but indicated the practical feasibility of the proposed approaches because generating training sets with human-labeled alignment may be difficult, and unsupervised approaches are certainly much more attractive if the performance can be acceptable.

Column (g) is the case very close to Baseline 2 in column (a) except utterance clustering was performed in addition and the tf-idf similarity was based on the clusters. We can see the utterance clusters did bring good improvements (columns (g) vs (a)) because it helped in the sparse/noisy word problem and considered the sequential smoothness assumption, but the entropy-based word filtering and reliability-propagated word-based matching used in column (b) was actually much stronger (columns (b) vs (g)). This verified the approaches proposed in sections 2.1 and 2.2 are useful. Column (h) is very close to column (e), structured SVM with supervised training, but using only the local features without the global features. We can see without the global features what the structured SVM could do was much less (columns (h) vs (e)), so learning from both local and global features was important here.

4. Conclusions

In this paper, we present the first known effort for the alignment between the spoken utterances in lectures and the slide content by combing a set of approaches including structured SVM considering the sparse/noisy word problem and the sequential smoothness assumption for slide presentation. Experimental results were very encouraging.

Utterance Transcriptions	(a) Baseline 2 Tf-idf Similarity	Proposed Approaches					Partial Tests	
		Unsupervised			Supervised		(g)	(h)
		(b) Word-based matching	(c) Structured SVM	(d) Score Integration	(e) Structured SVM	(f) Score Integration	(a) plus utterance clustering	(e) with Global Features Excluded
Manual	60.39%	74.74%	74.95%	76.58%	76.83%	77.16%	63.42%	73.41%
ASR	58.43%	69.50%	70.28%	72.86%	71.26%	73.15%	60.51%	68.97%

Table 1.：*Alignment accuracy for the proposed approaches compared to Baseline 2 based on tf-idf similarity.*

5. References

[1] T. Hayama, H. Nanba, and S. Kunifuji. Alignment between a technical paper and presentation sheets using a hidden markov model. In Proceeding of Active Media Technology, pages 102-106. IEEE, 2005.

[2] M.-Y. Kan. Slideseer: A digital library of aligned document and presentation pairs. In Proceedings of JCDL, pages 81-90. ACM, 2007.

[3] Bahrani, Bamdad, and Min-Yen Kan. "Multimodal alignment of scholarly documents and their presentations." Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries. ACM, 2013.

[4] Beamer, Brandon, and Roxana Girju. "Investigating automatic alignment methods for slide generation from academic papers." Proceedings of the Thirteenth Conference on Computational Natural Language Learning. Association for Computational Linguistics, 2009.

[5] Yih, Wen-Tau, and Christopher Meek. "Improving similarity measures for short segments of text." AAAI. Vol. 7. 2007.

[6] Hayama, Tessai, and Susumu Kunifuji. "Relevant piece of information extraction from presentation slide page for slide information retrieval system." Knowledge, Information, and Creativity Support Systems. Springer Berlin Heidelberg, 2011. 22-31.

[7] Vu, Thuy, Ai Ti Aw, and Min Zhang. "Feature-based method for document alignment in comparable news corpora." Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2009.

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

[9] T. Joachims. Learning to Align Sequences: A Maximum Margin Approach, Technical Report, August, 2003.

[10] 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.

[11] T. Joachims, Training Linear SVMs in Linear Time, Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD), 2006.

[12] T.Joachims, T. Finley, Chun-Nam Yu, Cutting-Plane Training of Structural SVMs, Machine Learning Journal, 77(1):27-59, 2009.

[13] Zhang, Shi-Xiong, and Mark JF Gales. "Structured Support Vector Machines for Noise Robust Continuous Speech Recognition." INTERSPEECH. 2011.

[14] Ramos, Juan. "Using tf-idf to determine word relevance in document queries."Proceedings of the First Instructional Conference on Machine Learning. 2003.

[15] Zhang, Wen, Taketoshi Yoshida, and Xijin Tang. "A comparative study of TF* IDF, LSI and multi-words for text classification." Expert Systems with Applicati-ons 38.3 (2011): 2758-2765.

[16] Tata, Sandeep, and Jignesh M. Patel. "Estimating the selectivity of tf-idf based cosine similarity predicates." ACM SIGMOD Record 36.2 (2007): 7-12.

[17] Lin, Jianhua. "Divergence measures based on the Shannon entropy." Information Theory, IEEE Transactions on 37.1 (1991): 145-151.

[18] S.-Y. Kong and L.-S. 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.

[19] Yeh, Ching-Feng, et al. "Recognition of highly imbalanced code-mixed bilingual speech with frame-level language detection based on blurred posteriorgram." Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on. IEEE, 2012.