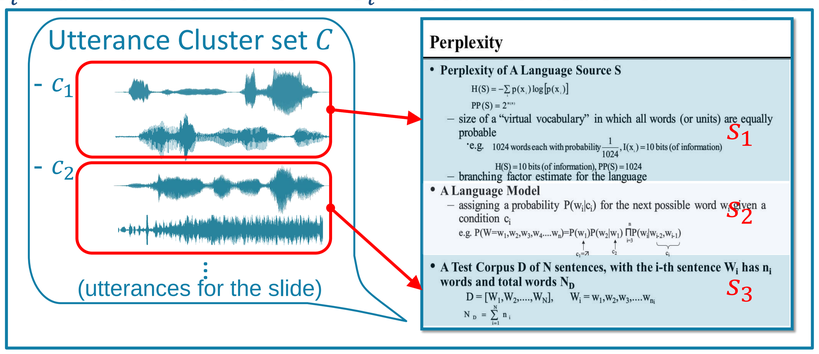
Alignment of Spoken Utterances with Slide Content for Easier Learning with Recorded Lectures using Structured Support Vector Machine (SVM)

Han Lu, Sheng-Syun Shen, Sz-Rung Shiang, Hung-Yi Lee, Lin-Shan Lee

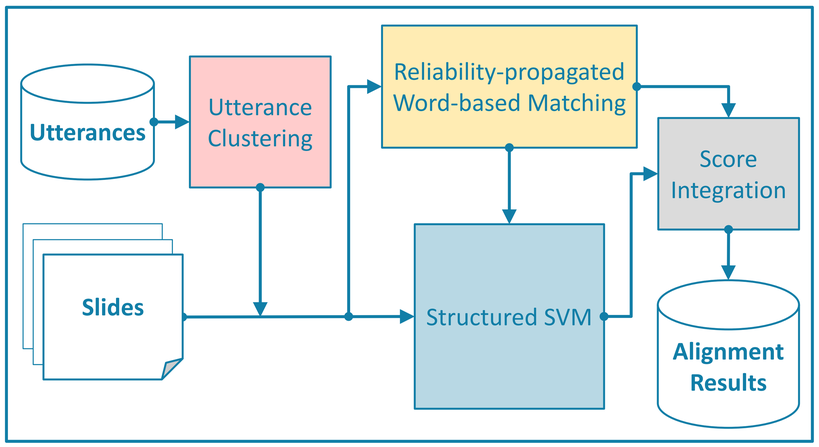
1. Introduction

■ Goal：
Automatically align the spoken utterances in recorded lectures with the slides.
Helpful in learning with MOOCs

■ Task Definition：
c_i：utterance clusters　　s_i：section based on subtitles



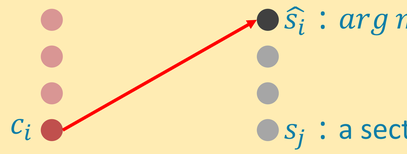
■ System Flowchart：



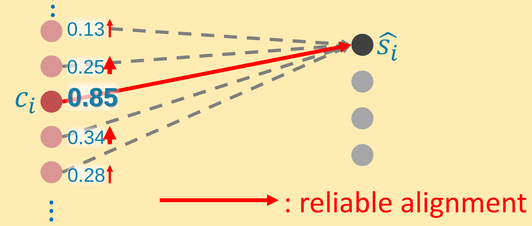
2. Proposed Approach

■ Utterance Clustering
To remedy the word scarcity and noisy word problems.

■ Reliability-Propagated Word-based Matching
- Word-based Matching
  Calculate lexical similarity based on tf-idf vectors.



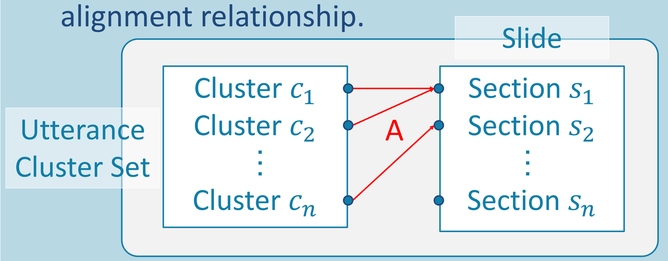
- Reliable Alignment
  $Sim(c_i, \hat{s}_i)$ much larger than $Sim(c_i, s_j)$ for other s_j
- Reliability-Propagated
  Score propagated from a reliable alignment to its neighbors.



■ Score Integration
Linearly combine the score obtained from Reliability-propagated Word-based Matching and Structured SVM. λ is tuned by development set.

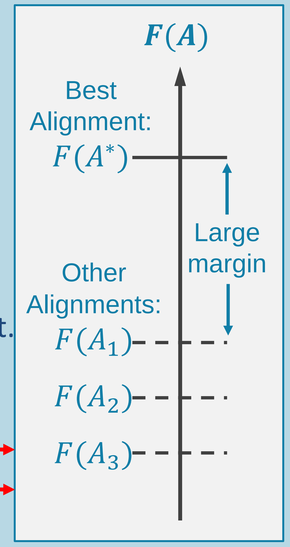
$$I(A) = F(A) + \lambda \sum_{i=1}^{|C|} Sim(c_i, s_{i\_align})$$

■ Structured SVM
- Consider the global alignment for a slide as a whole to optimize the alignment relationship.

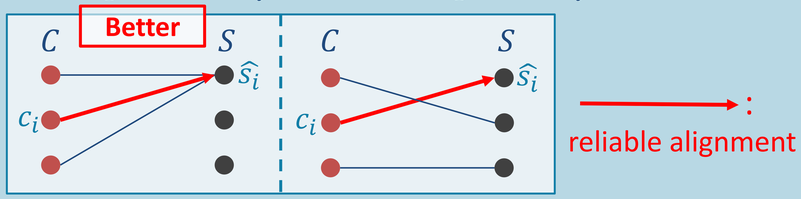


$$A = \{< c_i, \hat{s}_i > | i = 1,2,\dots,n\}$$

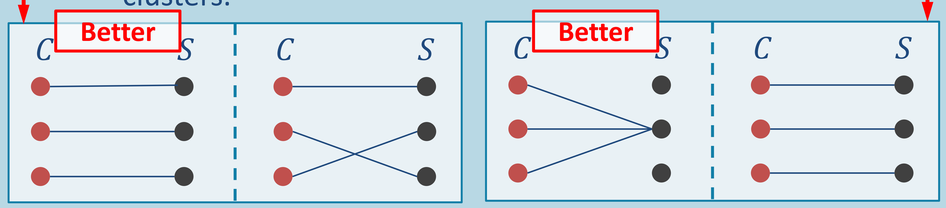
$\vec{v}(A)$：Characterizing quality of the alignment.
$\vec{w}$：To be learned.
$F(A) = \vec{w} \cdot \vec{v}(A)$：Evaluation Function.



- Training Goal：$max\ F(A)$ for training set.
- Testing Goal：$max\ F(A)$ for test slides.
- Local Feature Examples：
  1. Summation of lexical similarity：$\sum_{i=1}^{|C|} Sim(c_i, \hat{s}_i)$
  2. Reliability-Propagated method：If $< c_i, \hat{s}_i >$ is reliable, neighbors of c_i better also aligned to $\hat{s}_i$.



- Global Feature Examples：
  1. Number of crossed alignment
  2. Number of alignment transition：less transitions preferred
  3. Longer section should be explained with more utterance clusters.



- Structured SVM Training
  1. Supervised：Manually labeled training set.
  2. Unsupervised：Reliability-propagated Word-based Matching results as training set.

3. Experiment

■ Data set：A course in National Taiwan University.
■ Total length：45 hours with 193 corresponding slides.
(38 out of 193 used here.)
■ 4-fold Cross-validation：34 slides into 4 folds. 4 slides for development set.
For each trial, 3 fold for training and 1 fold for testing.
■ Conclusion：Structured SVM helped, and unsupervised performed very close to supervised after Score Integration.

Approaches		Accuracy
Baseline – Random		34.01%
Baseline – Tf-idf Similarity		58.43%
Word-based Matching		69.50%
Unsupervised	*Structured SVM*	70.28%
	Score Integration	72.86%
Supervised	*Structured SVM*	71.26%
	Score Integration	73.15%