INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Automatic Explanation Spot Estimation Method Targeted at Text and Figures in Lecture Slides

Shoko Tsujimura[1], *Kazumasa Yamamoto*[2], *Seiichi Nakagawa*[1,2]

[1]Department of Computer Science and Engineering, Toyohashi University of Technology, Japan
[2]Department of Computer Science, Chubu University, Japan

s155304@edu.tut.ac.jp, yamamoto@cs.chubu.ac.jp, nakagawa@tut.jp

Abstract

Because of the spread of the Internet in recent years, e-learning, which is a form of learning through the Internet, has been used in school education. Many lecture videos delivered at The Open University of Japan show lecturers and lecture slides alternately. In such video style, it is hard to understand where on the slide the lecturer is explaining. In this paper, we examined methods to automatically estimate spots where the lecturer explains on the slide using lecture speech and slide data. This technology is expected to help learners to study the lectures. For itemized text slides, using DTW with word embedding based distance, we obtained higher estimation accuracy than a previous work. For slides containing figures, we estimated explanation spots using image classification results and text in the charts. In addition, we modified the lecture browsing system to indicate estimation results on slides, and investigated the usefulness of indicating explanation spots by subjective evaluation with the system.
Index Terms: slide-speech alignment, lecture data, DTW

1. Introduction

Because of the spread of the Internet in recent years, e-learning, which is a form of learning through the Internet, has been used in school education. One of the advantages of e-learning is that learners can learn anytime and anywhere. However, in order to understand the contents of the lecture, they spend significant time looking through the entire contents of the lecture. Therefore, in order to be able to understand the contents of the lecture more efficiently, our research group created a lecture browsing system that added functions such as automatic speech transcription and automatic summarization to the output of the "EZ presentator" (Hitachi Advanced Digital Inc.) which synchronizes the video and slide switching time [1]. However, in this system, video, slide and speech transcription are separately presented, therefore there is a problem that it is hard to understand the correspondence between each part. Also, many lecture videos delivered at The Open University of Japan, which is one of the educational institutions engaged in e-learning, show lecturers and lecture slides alternately. In such video style, it is hard to understand where on the slide the lecturer is explaining. Therefore, it is expected for learners that indicating the spot where the lecturer explains on the slide will help to understand the contents of the lecture.

Some researchers have conducted studies to investigate whether indicating explanation spots affects the understanding of lectures. Takazawa et al. examined the learning effect by detecting the pointer from the lecture video and emphasizing the surrounding area so as to guide the learner's viewpoint [2]. As a result, it was found that emphasis processing contributed to the improvement of the understanding of the highlighted area.

Table 1: *Detail of lecture data*

Lecture name (in this paper)	Lecture A	Lecture B
Lecture ID	L11M0010	L11M0030
Lecture name	Advanced computer application 2	
Total time	01:07:56	01:05:49
Word accuracy(%)	54.9	70.6

Ando et al. examined the difference in the learning effect depending on the presence or absence of the pointer, and conducted two types of tests, a memory test to answer terms and an understanding test of an essay-type [3]. As the result, when images were included in the slide, it turned out that the correct answer rate is higher when using pointer system in both tests. Based on these research results, the effectiveness of indicating explanation spots can be considered.

For lecture retrieval, researches for aligning lecture slides with speech recognition results or articles are relatively common [4, 5, 6, 7]. However, rarely are studies undertaken to align individual contents in the slide with utterances, as dealt with in this paper. Matsuda et al. extracted keywords from slides and speech recognition results on the itemized text slide and estimated the explanation spots using the number of keyword matches [8]. However, utterances that are different from slide notation but have similar expressions cannot be estimated well since this method estimates explanation spots only depending on the number of exact matches of keywords. Lu et al. divided speech into clusters and used features such as tf-idf to align speech clusters with places in the slide [9]. However, when the same keyword appears in multiple utterances or places in the slide, aligning may not be successful. Marutani et al. studied to extract objects indicated by the lecturer's deictic gesture using information about lecturer's gestures and direction of the lecturer's body [10]. In this method, however, the lecturer's gesture needs to be recorded and some beacons must be attached to the lecturer's shoulders and hands, so it is difficult to realize.

In this paper, we examined methods to automatically estimate spots where the lecturer explains on the slide using lecture speech and slide data without using lecturer gestures. In addition, we modified the lecture browsing system to indicate estimation results on slides, and investigated the usefulness of indicating explanation spots by subjective evaluation with the system.

2. Overview of lecture data

In this research, we use two lectures to which slide data belongs, among the data included in the Corpus of spoken Japanese Lecture Contents (CJLC) monitor version [11]. Table 1 shows details of lecture data used in this research. Although the subject names of lectures are the same, lecture contents and the speaker are different. Slides used in each lecture are written in Japanese, and lecturers explain in Japanese. Each lecture is hereinafter referred to as "Lecture A" and "Lecture B".

http://dx.doi.org/10.21437/Interspeech.2017-750

2.1. Lecture speech recognition

ASR was carried out using LVCSR system for recorded speech data of Lecture A and Lecture B. As the decoder, we used SPOJUS++ [12] WFST version. The feature vector consists of 40 FBANKs along with their first and second derivatives. DNN-HMM was used for the acoustic model, and the number of input frames to DNN is 11 connecting the 5 frames before and after the current frame. The structure of DNN is 1,320 inputs (120 features $\times$ 11 frames), 8 hidden layers with 2,048 hidden units, and 3,015 outputs based on the number of states of HMM. The HMM models tied-state triphones. The male academic lecture speech (about 122 hours) in Corpus of Spontaneous Japanese (CSJ) [13] was used for training DNN-HMM. A tri-gram based language model was trained on 970 lectures of CSJ (vocabulary size of 20k words) morphologically analyzed by MeCab [14] with ipadic [15].

The lecture speech data was recorded using a hand microphone. It was automatically divided into utterance units by silent sections of 200 ms, and used for speech recognition. Word accuracy is shown in the last row of Table 1.

3. Definition of segment unit

In this research, we divide a lecture slide into smaller units for each group of meanings, and estimate and indicate them as the spot where the lecturer explains (hereinafter referred to as "segment"). In this section, we describe how to divide slide into segments. The process described below can almost automatically divide Microsoft PowerPoint slide data (.pptx format) written in OpenXML format [1].

3.1. Segmentation for itemized text

For itemized text, the same units as in [8] are set as segments. Specifically, each itemized structure is divided into segments according to the definition below (Figure 1):

(1) When bullets or line feeds are made, it is assumed that the contents has changed, and it is set as the first segmentation candidate (automatic line feeds that occur due to too long line are not candidates for segmentation).

(2) If there is only one line in the second level of itemization on the inside of the first level of itemization, it is considered that the contents of the first level cover the contents of the second level, so the second level is included in the same segment as the first level.

(3) If there are two or more itemized lines in the second level of itemization on the inside of the first level of itemization, it is considered that each itemization has independent content, so it is made as a separate segment.

(4) In the case of a structure with three or more levels of itemization, assuming that the contents of the third level are complements of the second level, in order to prevent excessive division, all three or more levels of itemization are included in the same segment as the second level.

3.2. Segmentation for text boxes

A text box basically defines one line as one segment. However, if they are semantically collective, even multiple lines are defined as one segment.

[1] We converted to .pptx format using Microsoft PowerPoint compatibility mode because slide data distributed in the CJLC monitor version is in .ppt format.

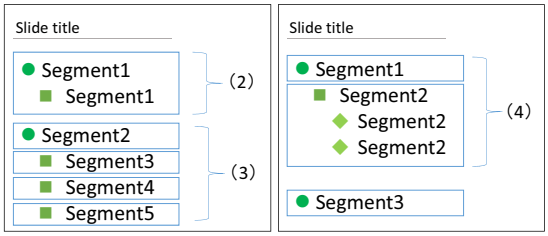

Figure 1: *Example of segmentation for itemized text slide*

3.3. Segmentation for charts and images

Charts and images are defined as one segment for each chart or image not depending on the format (whether a chart is created by Excel or attached as an image).

4. Automatic explanation spot estimation for itemized text slides

In this section, we describe the method for estimating explanation spots for itemized text slides and the result. Itemized text slides shall be divided into segments as described in Section 3.1 for each slide. Because the slide change time is known by the function of the EZ presentator, utterances spoken in each slide are known. The division into the utterance unit has also already been performed.

4.1. Estimation method

Estimation of the explanation spot is performed by using keywords extracted from speech recognition result of each utterance and text data in each segment. Keywords are nouns, adverbs and original forms of adjectives and verbs, and they are obtained by morphologically analyzing each segment and utterance. Division into morpheme units is done by MeCab with unidic [16].

In the case of itemized text slides, the lecturer is likely to explain the contents of each slide in order from the top to the bottom. Therefore, in this research, by doing DTW using all segments and utterances in one slide, we make an alignment while considering the time series. The local distance of DTW between one utterance and one segment is calculated by the following method using such as similarity between keywords and the appearance frequency of keywords.

U : keyword list of an utterance
S : keyword list of a segment
T : threshold of similarity
df_i : number of segments in which keyword i appears in the slide

$s\_sum = 0, m\_sum = 0;$
For $i = 1$ to $|U|$
  $maxsim = 0;$
  For $j = 1$ to $|S|$
    $sim \leftarrow$ similarity between $U(i)$ and $S(j)$
    If $sim > maxsim$ then $maxsim = sim, key = S(j);$
  If $maxsim < T$ then $maxsim = 0, m = 0;$
    otherwise $m = 1;$
  $f = maxsim \times \frac{1}{df_{key}};$
  $s\_sum = s\_sum + f;$
  $m\_sum = m\_sum + m;$
$LD = 1 - \frac{s\_sum}{|U|} \times \frac{m\_sum}{|S|}$ and LD is used as local distance.

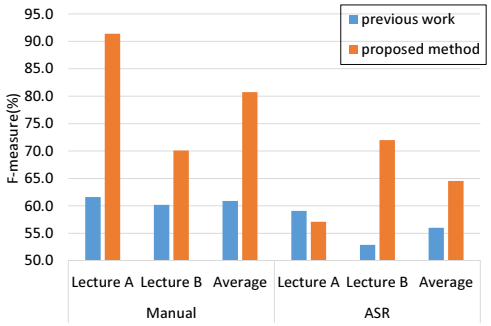

Figure 2: *Estimation result for itemized text slides*

For similarity calculation between keywords, we use word2vec [17] of word embedding and use cosine similarity between word vectors as the similarity between keywords. In this research, we use the word2vec model trained by Skip-gram on Japanese Wikipedia data (vocabulary size of 420 million words) [18]. The threshold value $T = 0.7$ is used, which is the highest accuracy in preliminary experiments.

Utterances not including keywords are excluded from DTW and DTW is performed only for utterances and segments including keywords. Then, after aligning the utterance with the segment, utterances not including keywords are set to the same explanation spot as the segment estimated by preceding utterance, so that all utterances are aligned with one or more segments.

4.2. Estimation result

Explanation spots were estimated by the proposed method and compared with Matsuda's method [8]. The estimation result was evaluated by F-measure. The comparison result is shown in Figure 2. In Figure 2, "Manual" means the result using manual transcription, and "ASR" means the result when speech recognition result is used.

In Matsuda's method [8], keywords in one utterance are compared with keywords in each segment in the slide, and a segment with the largest number of matching keywords is selected as an estimation explanation spot. The main difference from the proposed method is that DTW is not used and explanation spot is estimated only by the number of matching keywords. From Figure 2, it can be seen that the estimation accuracy of the proposed method is higher than Matsuda's method except for the "ASR" of Lecture A. Therefore, it can be said that the proposed method can more accurately estimate explanation spots than Matsuda's method.

5. Automatic explanation spot estimation for figures slides

In this section, we describe the method of estimating explanation spots for slides that cannot use the method in Section 4, which includes that such as a block diagram composed of text boxes, charts and images (e.g. Figure 3), and the results.

5.1. Keyword definition for charts and images

For slides including figures, estimation of explanation spots is performed by using keywords extracted from the speech recognition result of each utterance and text data in each segment. However, for text boxes and charts created in Excel format text can be acquired from the data, but cannot for images. Therefore, in this paper, we consider image classification and using information on the kind of image as keywords.

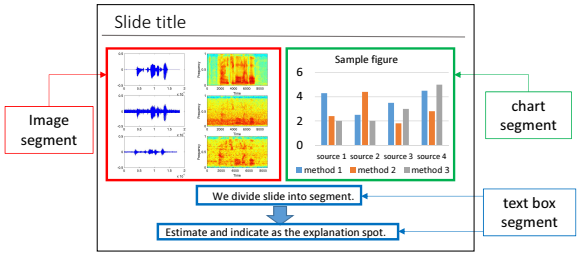

Figure 3: *Example of figure slide*

For image classification, we use SVM and the neural network model (it became the top in ILSVRC2012 [19]) which is already trained as an image classification model of Caffe which is a framework of deep learning. Image classification is performed by using 4,096 dimensional feature vector obtained from the sixth hidden layer that is the middle layer of the model as SVM inputs and classifying into four categories (table, graph, flowchart, and others) that are frequently used in lectures. For SVM training, a total of 400 images (100 images for each category) obtained from Google image search results were used. According to the image classification result by the SVM, "table" in the case of the table, "graph, chart" in the case of the graph, "flowchart" in the case of the flowchart and "photograph, image, drawing, picture" in the case of the others were defined as keywords of the image segment, respectively.

When two or more charts and images are included in one slide, the lecturer may talk about information based on the position of an object (such as a chart or an image) included in the slide, such as "the figure on the right". Therefore, when two or more objects are included in one slide, the position information is added to the keyword of the segment. The position information of the object included in the slide can be acquired by using the data from PowerPoint. By comparing the acquired (x, y) coordinate values with the center coordinates of the slide, we added corresponding position information among "right, left, up, down, upper right, lower right, upper left, lower left, center, and middle" to the keywords of the segment.

5.2. Estimation method

Even for slides including figures, estimation of explanation spots is performed using keywords extracted from the speech recognition result of each utterance and keywords defined in Section 5.1 in each segment. The estimation method is as follows. For the calculation of similarity between keywords, the model of word2vec in Section 4.1 is used and the value of threshold T is also 0.7.

S : number of segments in a slide
U_K : keyword list of an utterance
S_K : keyword list of a segment
U_C : list of compound words in an utterance. This list is created by describing a compound word that is organized into one word and morpheme number of the composed word when keywords appear consecutively while extracting keywords by morphological analysis. In the case of a slide including figures, the lecturer tends to speak the contents written on the slide as it is, so we use compound words.
S_C : list of compound words in a segment. This list is created in the same way as U_C.
T : threshold of similarity

For $i = 1$ to S
  $s\_sum = 0, m\_sum = 0;$
  For $j = 1$ to $|U\_K|$
    $maxsim = 0;$
    For $k = 1$ to $|S\_K|$
      $sim \leftarrow$ similarity between $U\_K(i)$ and $S\_K(j)$
      If $sim > maxsim$ then $maxsim = sim;$
    If $maxsim < T$ then $maxsim = 0;$
                otherwise $maxsim = 1;$
    $s\_sum = s\_sum + maxsim;$
  $f = \frac{s\_sum}{|U\_K|}$
  For $n = 1$ to $|U\_C|$
    $max\_c = 0;$
    For $m = 1$ to $|S\_C|$
      If $U\_C(n) = S\_C(m)$
        $c \leftarrow$ the number of morphemes
            of the compound word $S\_C(m)$
        If $c > max\_c$ then $max\_c = c;$
  $SCORE = f \times max\_c;$
The segments with the highest $SCORE$ is selected as estimated explanation spots.

5.3. Estimation result

Using the method in Section 5.2, we estimated explanation spots for Lecture A and Lecture B. As a result, the F-measure is 40.0% in "Manual" and 34.1% in "ASR" in average, which is low estimation accuracy. Also, the image classification accuracy was as low as 45.5% with 11 images used in Lecture A and Lecture B. However, in the lectures, the keywords defined based on the classification result rarely appeared in the utterance, so it is considered that the image classification accuracy did not affect much the estimation accuracy of the explanation spots. Instead of that, because the lecturers used pointers, the instruction words such as "this" are frequently used. Therefore, it seems that the keywords did not generally appear during the utterances, and estimation was more difficult.

6. Subjective evaluation experiments

We conducted an evaluation experiment of the proposed method using lecture data of about 15 minutes, two slides from Lecture A and six slides from Lecture B (five itemized text slides and three figures slides). The slides used for the evaluation experiment were selected based on the fact that they handled semantically contiguous contents so that the subjects could easily learn, and the balance of itemized text slides and figures slides. We prepared three systems: a system which does not indicate the explanation spots on the slide (original) [1], a system which indicates manually correct answer spots (oracle), and a system which indicates estimation results by the method described in Section 4 and 5 with ASR result. We asked subjects to attend the same 15 minute lecture on each system and to answer questionnaire asking for easy-understand, etc. The subjects answered the questionnaire by using the 7 point Likert scale (from "7: I agree very much" to "1: I do not think so at all"). The subjects are ten students in the first and second year of Master's degree, and they are students who have not taken "Advanced Speech and Language Processing" currently being offered at our university, which deals with contents almost similar to Lecture A and Lecture B. In order to eliminate the order effect, the order of using the three systems was randomized for each subject.

After using three systems, we asked subjects to answer the following three questions for each system.

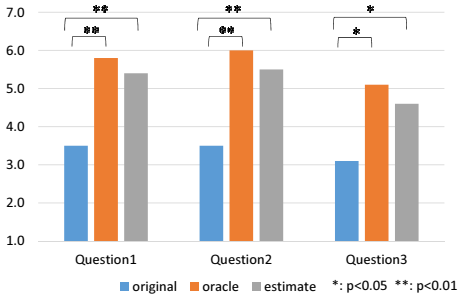

Figure 4: *Average of each question*

Q1 Was this system easy-to-understand when learning?
Q2 Do you think this system will be useful for learning?
Q3 Do you want to learn using this system?

First, we calculated the average value of answers to each question for each system. The results are shown in Figure 4. From Figure 4, it can be seen that the system that indicates explanation spots has higher evaluation results than the system that does not indicate in all questions.

Next, in order to evaluate the answers of each question between the systems, we performed significance test (two-sided t-test). As a result, significant differences were observed between original and estimate as well as between original and oracle in all questions. Therefore, it is thought that it is possible to better understand lectures by using a system that indicates explanation spots rather than a system that does not indicate. Moreover, there was no significant difference between oracle and estimate even at the significance level of 10% in all questions. Therefore, even using the system that indicates the explanation spots estimated by the proposed method, there was no large difference of easy-to-understand when using the oracle system. As a result of an interview on the answers, almost half of the subjects did not perceive the difference between oracle and estimate, despite the low estimation accuracy of the estimate system for figures.

7. Conclusions

We proposed methods to automatically estimate spots where the lecturer explains on the slide using lecture speech and slide data. As a result, it was possible to obtain higher estimation accuracy on the itemized text slides. However, for figures slides, the F-measure is as low as 30%, which is the result where improvement in accuracy is required. Also, we modified the lecture browsing system to indicate explanation spots estimated by the proposed methods on slides and conducted subjective evaluation experiments. As a result, the system that indicates explanation spots was significantly more highly evaluated than the system that does not indicate, and the system that indicates estimated explanation spots was found to be also useful.

In the future, we are considering increasing the lecture data and evaluating the proposed method using lectures not using a pointer. In addition, since the proposed method for itemized text slides estimates using DTW, it cannot correctly estimate when returning to the already described segment or explaining multiple spots with one utterance. Therefore, by using the proposed method in combination with other methods, we would like to make it possible to estimate according to actual lecture styles. Moreover, it is necessary to increase the number of subjects and to examine the effectiveness of indicating explanation spots in more detail by examining the learning effect when learners learn using the system that indicates explanation spots and the system that does not indicate, respectively.

8. References

[1] S. Togashi and S. Nakagawa, "A browsing system for classroom lecture speech," *INTERSPEECH 2008*, pp. 2803–2806, 2008.

[2] T. Takazawa, S. Fukuma, Y. Oizumi, A. Okuda, and T. Sakurai, "A pointing system for effective distant learning and its evaluation," *IEICE General Conference, D-15-15*, p. 168, 2007, (in Japanese).

[3] M. Ando and M. Ueno, "Effect of pointer presentation on multimedia e-learning materials," *Proceeding of World Conference on Educational Multimedia, Hypermedia & Telecommunications*, pp. 5549–5559, 2008.

[4] S. Nakazawa, K. Satoh, and A. Okumura, "Alignment of lecture speech data and presentation documents for lecture search," *IPSJ SIG Technical Report*, vol. 2004-SLP-55, no. 12, pp. 65 – 70, 2005, (in Japanese).

[5] S. Tanaka, T. Tezuka, A. Aoyama, F. Kimura, and A. Maeda, "Slide retrieval technique using features of figures," *Proceedings of the International MultiConference of Engineers and Computer Scientists 2013*, vol. 1, pp. 1–6, 2013.

[6] M.-Y. Kan, "Slideseer: A digital library of aligned document and presentation pairs," *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, pp. 81–90, 2007.

[7] B. Bahrani and M.-Y. Kan, "Multimodal alignment of scholarly documents and their presentations," *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries*, pp. 281–284, 2013.

[8] E. Matsuda, Y. Horiuchi, and S. Kuroiwa, "Estimation of explanation spot in lecture slide using speech information for OCW," *IPSJ SIG Technical Report*, vol. 2011-CLE-4, no. 6, pp. 1–8, 2011, (in Japanese).

[9] H. Lu, S. syun Shen, S.-R. Shiang, H. yi Lee, and L. shan Lee, "Alignment of spoken utterances with slide content for easier learning with recorded lectures using structured support vector machine (SVM)," *INTERSPEECH 2014*, pp. 1473–1477, 2014.

[10] T. Marutani, S. Nishiguchi, K. Kakusho, and M. Minoh, "Making a lecture content with deictic information about indicated objects in lecture materials," *AEARU Workshop on Network Education*, pp. 70–75, 2005.

[11] "Corpus of spoken Japanese Lecture Contents (CJLC)," http://www.slp.cs.tut.ac.jp/CJLC/ (Accessed 1 June 2017).

[12] Y. Fujii, K. Yamamoto, and S. Nakagawa, "Large vocabulary speech recognition system: SPOJUS++," *International Conference MUSP*, pp. 110–118, 2011.

[13] "Corpus of Spontaneous Japanese (CSJ)," http://pj.ninjal.ac.jp/corpus_center/csj/ (Accessed 1 June 2017).

[14] "MeCab: Yet Another Part-of-Speech and Morphological Analyzer," http://taku910.github.io/mecab/ (Accessed 1 June 2017).

[15] "IPAdic legacy," https://osdn.net/projects/ipadic/ (Accessed 1 June 2017).

[16] "UniDic Project Top Page," https://osdn.net/projects/unidic/ (Accessed 1 June 2017).

[17] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *Proceedings of Workshop at ICLR*, pp. 1–12, 2013.

[18] "Wikipedia: Database download," https://jp.wikipedia.org/wiki/Wikipedia:Database_download (Accessed 1 June 2017).

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Advances in 25th Neural Information Processing Systems*, pp. 1106–1114, 2012.