Developing Corpus of Lecture Utterances Aligned to Slide Components

Ryo Minamiguchi and **Masatoshi Tsuchiya**
Department of Computer Science and Engineering
Toyohashi University of Technology
1–1 Hibarigaoka, Tempaku-cho, Toyohashi, Aichi, Japan
minamiguchi@is.cs.tut.ac.jp and tsuchiya@cs.tut.ac.jp

Abstract

The approach which formulates the automatic text summarization as a maximum coverage problem with knapsack constraint over a set of textual units and a set of weighted conceptual units is promising. However, it is quite important and difficult to determine the appropriate granularity of conceptual units for this formulation. In order to resolve this problem, we are examining to use components of presentation slides as conceptual units to generate a summary of lecture utterances, instead of other possible conceptual units like base noun phrases or important nouns. This paper explains our developing corpus designed to evaluate our proposing approach, which consists of presentation slides and lecture utterances aligned to presentation slide components.

1 Introduction

Automatic text summarization is one of the tasks that have long been studied in natural language processing area. One of well-known approaches for automatic text summarization is an extractive method which picks important textual units (e.g. sentences) from given documents (Kupiec et al., 1995; Goldstein et al., 2000; Radev et al., 2000).

(Filatova and Hatzivassiloglou, 2004) introduced *conceptual units* to represent meaning components, and formulated the extractive method of text summarization as a maximum coverage problem with knapsack constraint (henceforth, denoted as MCKP). Suppose a finite set T of textual units which means whole given documents, and a finite set C of conceptual units which represents whole information described by T. In this representation, a textual unit may describe one or more conceptual units, and an information overlap between picked textual units is considered as a redundant conceptual unit(s) which is described by plural textual units. In other words, the meaning of each textual unit is regarded as a subset of C, and the extractive method of text summarization is defined as a problem to find a subset of T which satisfies the constraint of its total length and describes as many conceptual units as possible. Various methods including greedy algorithm (Filatova and Hatzivassiloglou, 2004), stack decoding (Yih et al., 2007) and linear programming solver (Takamura and Okumura, 2009) were employed to solve text summarization in this representation.

This representation provides a concrete and concise formulation of text summarization, however, a big problem still remains: the appropriate granularity of conceptual units. (Hovy et al., 2006) proposed to use basic elements as conceptual units, which are dependency subtrees obtained by trimming dependency trees. (Takamura and Okumura, 2009) proposed to use weighted content words as conceptual units, whose weights reflect their importance. Although these possible conceptual units treat linguistic clues of original documents, they do not represent the intuition of the writer (or the speaker) of the original documents.

In order to resolve this problem, we are examining to extract dependency structure between primitive objects such as texts, pictures, lines and basic diagrams, and to use these objects as conceptual units when generating a summary of lecture utterances. We think that this approach has two advantages than the previous approach of conceptual units. The first is that terminology and character formatting of these objects may reflect the intuition of the lecturer about his/her talk, because these objects are selected

Proceedings of the 12th Workshop on Asian Language Resources,
pages 30–37, Osaka, Japan, December 12 2016.

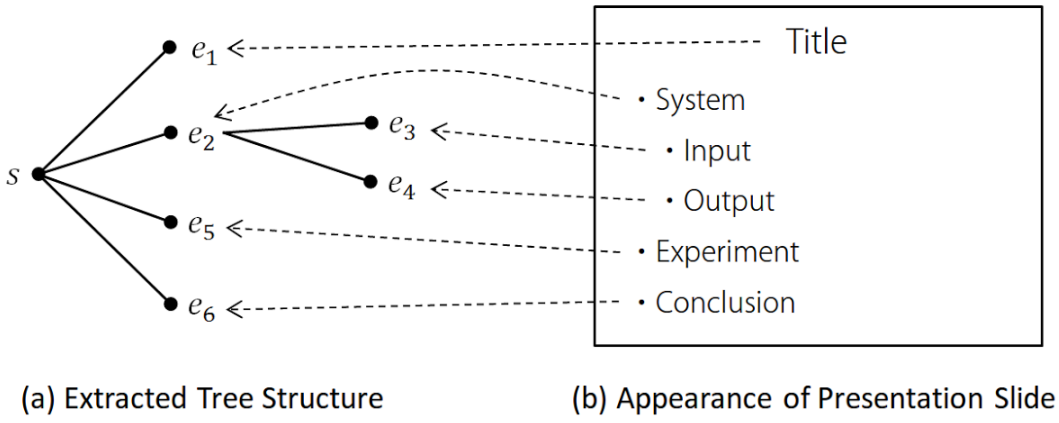

Figure 1: A presentation slide example

and located by him/herself. For example, he/she will use either a larger point font or a bold style font, to represent an important part of his/her talk. The second is that this approach naturally introduces multi-level granularity of conceptual units because our using method proposed by (Hayama et al., 2008) extracts relationship between objects as a tree structure. When multi-level granularity of conceptual units is available, the remaining problem to decide appropriate granularity of conceptual units can be considered as a simple optimization problem.

This paper explains our developing corpus which consists of lecture utterances, presentation slides, and their alignment information. We think that this corpus will give a foundation to evaluate our assumption about conceptual units.

2 Structure of Presentation Slide

Generally speaking, a presentation slide consists of one or more primitive objects, such as texts, pictures, lines and basic diagrams. We call these primitive objects as *slide components* in this paper. Slide components are carefully located in a presentation slide by its author, taking his/her presentation speech procedure into consideration. Thus, from the human view point, a dependency structure between slide components represented by either their relative positional relationship or basic diagrams including an arrow sign emerges.

Unfortunately, it is necessary to extract the dependency structure between slide components, because it is not explicitly represented in the slide data itself. We employ the method proposed by (Hayama et al., 2008), which uses relative positional relationship between slide components to extract dependency structure. Figure 1 shows an example of presentation slide designed in the traditional style and the dependency structure extracted from it. The root node represents the slide s itself. The root node has children including the headline e_1 of the slide, the first-layer bulleted ed text snippets e_2, e_5, and e_6. And more, the node e_2 has the second-layer bulleted text snippets e_3 and e_4 as the children of e_2.

It is true that our using method cannot extract structures from all styles of presentation slides. In the modernized style introduced in (Alley et al., 2005), basic diagrams play more important role to represent relationships between slide components than ones in the traditional style. Because our using method uses relative positional relationship between slide components as key clues and does not handle basic diagrams, it faces limitation against the modernized style slides. However, the dependency structure between slide components still exists in the modernized style, and an improved structure extraction method will resolve this limitation.

3 Alignment between Slide Components and Lecture Utterances

This section describes the detail of our corpus design.

Table 1: Statistics of CJLC

# of speakers	15
# of courses	26
# of lectures	89
Duration	3,780 min.

Table 2: Age of Speakers, their teaching history and number of their courses

	minimum	average	maximum
Age of speakers	31	41.5	58
Teaching history	2	14.2	30
# of courses	2	4.2	7

3.1 Corpus

Corpus of spoken Japanese Lecture Contents (henceforth, denoted as CJLC) developed by (Tsuchiya et al., 2008) is used as the main target of this research. It is designed as the fundamental basis of researches including large vocabulary continuous speech recognition and extraction of important sentences against lecture contents, and consists of speech, transcriptions, and presentation slides that were collected in real university classroom lectures. Thus, we think that the design objective of CJLC matches well our research.

CJLC is formally defined as a set of classroom lecture data, and each data consists of following 5 items:

1. a lecture speech recorded with several microphones,
2. its synchronized transcription,
3. a presentation slide data (Microsoft PowerPoint formed),
4. a timetable of slide show, and
5. a list of important utterances.

Table 1 shows the statistics of CJLC. Generally speaking, a course of CJLC is a series of one or more lectures. All speeches of CJLC were transcribed by human annotators. Table 2 shows the distribution of 15 speakers recorded in CJLC. A lecture speech data and its synchronized transcriptions are provided for all lectures, but a presentation slide data, a timetable of slide show and a list of important utterances are not attached to all lectures.

Note that each speech of CJLC was automatically segmented into utterances using the amplitude of the speech signal described in (Kitaoka et al., 2006; Otsu, 1979), and that their segmentation do not match to sentence boundaries for spontaneous speech proposed by (Takanashi et al., 2003). Although it means that automatically segmented utterances of CJLC are not sentential units from the view point of their senses, automatically segmented ones are referred as textual units, for two reasons. The first reason is that automatic detection methods of sentence boundary against spontaneous speech were proposed by (Shitaoka et al., 2004; Akita et al., 2006), however, they do not achieve sufficient performance when results of automatic speech recognition contain many errors. The second reason is to keep compatibility with important utterance extraction information of CJLC.

3.2 Alignment Labels

Four labels are introduced to represent alignment information between textual units and slide components. First of all, Label **I** and Label **O** are introduced to distinguish whether textual units correspond

	Content of utterance	Aligned slide component
t_a	Generally, a computer system has two kinds of operation devices, such as an input device, an output device.	e_2
	⋮	
t_b	Typical input devices are a keyboard and a mouse, and typical output devices are a monitor and a speaker.	e_2
	⋮	
t_c	In this stage, the monitor displayed the photos and the videos for the experiment.	e_4/e_5

Figure 2: Example of alignment label with slide components

to any slide components or not. Label **B** and Label **E** are introduced to resolve mismatch between automatic power-based boundary and sentence boundary. The following is more detailed descriptions of these labels.

- Label **I** means that its labeled textual unit is either an utterance or a part of an utterance to explain a slide component. An explanation may be carried by either a same content word, a synonym, a hypernym, a hyponym, a paraphrase, an expression to instantiate a general case shown by the slide component into a specific case, or an expression to abstract a specific case shown by the slide component into a general case. When the textual unit t_i explains the slide component c_j, a pair of Label **I** and the sequence number j is assigned to t_i.

- Label **B** means that its labeled textual unit belongs to the succeeding textual unit from the view point of sentence boundary, only when the succeeding unit has either Label **I** or Label **B**. In other words, the textual unit which has Label **B** is a former part of a sentence, which must contain one or more textual units which have Label **I**.

- Label **E** is the opposite label of Label **B**, and means that its labeled textual unit belongs to the preceding textual unit from the view point of sentence boundary only when the preceding unit has either Label **I** or Label **E**. In other words, the textual unit which has Label **E** is a latter part of a sentence, which must contain one or more textual units which have Label **I**.

- Label **O** means that its labeled textual unit are not related to any slide components.

The alignment label system described in the above can represent the case that one or more textual units explain a slide component. It, however, involves difficulty for the case that a single textual unit explains multiple slide components.

In order to conquer this difficulty, this case is divided into three sub cases, and procedures to select an appropriate slide component are prepared. Figure 2 shows example of alignment label with slide components in three sub cases.

The first sub case is that a parent-child relationship exists between the two slide components explained by a single textual unit. Suppose that the slide component e_2 and the slide component e_3 of Figure 1 are explained by the single textual unit t_a. In this corpus, the parent node e_2 is selected as the label of the textual unit t_a. The second sub case is that a sibling relationship exists between the two slide components explained by a single textual unit, and that two slide components share the same parent node. The example of the second sub case is that the slide component e_3 and the slide component e_4 of Figure 1 are explained by the single textual unit t_b. In this corpus, the parent node e_2 shared by the explained nodes e_3 and e_4 is selected as the label of the textual unit t_b. The last sub case is the rest of the above sub cases. For example, suppose that the slide component e_4 and the slide component e_5 are explained by the single textual unit t_c. In order to resolve the last sub case, both e_4 and e_5 are recorded in parallel as the label of t_c while annotation work. Because the last sub case is rare, for the following

analysis of this paper, the preceding node e_4 is referred as the label of t_c and the succeeding node e_5 is ignored.

The alignment manual for annotators reflects the descriptions of labels explained in the above. The following is the abstract of the manual.

1. The supervisor supplies a set of textual units and a set slide components to the annotator.
2. The annotator is requested to find all kind of explanations and to assign all Label **I** in the given set. When a single textual unit explains multiple slide components, the annotator must select an appropriate node in compliance with the procedures described in the above.
3. After assignment of Label **I**, the annotator is requested to find all Label **B** and Label **E** in the given set. In other words, the annotator must find sentence boundaries around textual units labeled as Label **I**.
4. After that, Label **O** is assigned to all remaining textual units.

3.3 Annotation Results

Two annotators[1], who are master course students of the department of computer science, are employed for the annotation work of the corpus. Table 3 shows their annotation results. Each lecture has a lecture ID (for example, *L11M0011*) which is composed of four parts: its first part is a letter *L* which means a first letter of lecture, its second part is a two didit number *11* which identifies a anonymized speaker, its third part is a letter *M* which means a gender of a speaker, and its last part is a four digit number *0011* which distinguishes a lecture. Furthermore, the last four digit number is composed of two sub parts: its first sub part is a three digit number *001* which means a course, and its second sub part is a one digit number *1* means the sequence number of the specified lecture in the course. In order to measure agreement of two human annotators' results, the following κ statistics (Chklovski and Mihalcea, 2003; Ng et al., 1999) is widely used.

$$\kappa = \frac{P_a - P_e}{1 - P_e} \tag{1}$$

Here, P_a denotes the empirical agreement ratio between two human annotators, while P_e denotes the probability of agreement by chance.

The annotation label system of our corpus is two layered: the first layer labels, such as Label I, Label B, Label E and Label O, represent whether their labeled textual units are related to slide components or not, and the second layer, which consists of sequence numbers of Label I, represents explanation relationships between textual units and slide components. In order to measure fairly agreements of this two layered label system, two kinds of granularity are introduced when computing κ statistics. When computing κ statistics for coarse granularity to measure the agreement of the first layer labels, the empirical agreement ratio P_a is defined as the following equation.

$$P_a = \frac{\sum_{X=\{\mathrm{I,B,E,O}\}} a(X)}{|T|} \tag{2}$$

$a(X)$ is the number of textual units which two human annotators give the same label X, and $|T|$ is the number of textual units. The probability of agreement by chance P_e is calculated as follows:

$$P_e = \sum_{X=\{\mathrm{I,B,E,O}\}} P^2(X), \tag{3}$$

where $P(X)$ is the label occurrence probability. When maximum likelihood estimation is employed, $P(X)$ is defined as follows:

$$P(X) = \frac{f(X)}{|T|}, \tag{4}$$

[1] An annotators is one of the authors.

Table 3: Result of Manual Annotation

Lecture ID	# of slides	# of slide components	# of utterances	# of labels				κ statistics	
				I	B	E	O	coarse	fine
L11M0010	21	370	742	578	4	26	134	0.68	0.61
L11M0011	29	431	704	584	11	14	95	0.58	0.72
L11M0012	12	276	811	546	2	5	258	0.83	0.65
L11M0030	58	822	680	414	41	57	168	0.92	0.75
L11M0050	22	159	2362	1280	39	81	962	0.68	0.6
L11M0064	27	469	1110	559	51	58	442	0.69	0.72

where $f(X)$ is the number of textual units to which Label X is assigned.

When computing κ statistics for fine granularity to measure the agreement of the second layer labels, which means the agreement of sequence numbers, the empirical agreement ratio P'_a is defined as the following equation.

$$P'_a = \frac{\sum_{j=1}^{|C|} a(c_j)}{f(\mathrm{I})}, \tag{5}$$

where $|C|$ is the number of slide components, and $a(c_j)$ is the number of textual units which are associated to the same slide component c_j. When the probability of agreement by chance is calculated for fine granularity, as already described in Equation 3, the probability which a slide component c is assigned to textual units by a human annotator is required. When uniform distribution is assumed in order to avoid zero frequency problem, it is defined as follows:

$$P(c) = \frac{1}{|C|} \tag{6}$$

The larger the κ statistics, the more reliable the results of the human annotators. (Carletta, 1996) reported that $\kappa > 0.8$ means good reliability, while $0.67 < \kappa < 0.8$ means that tentative conclusions can be drawn. According to his criteria, when measuring the agreement of two human annotators for coarse granularity, the reliability level of 2 lectures is good, the reliability level of three lectures is tentative, and the rest lecture, L11M0011, is not reliable. Its presentation slide contains many figures, and our using method to extract slide components from the presentation slide has the limitation to handle figures as already described in Section 2. We think that this limitation causes the inagreement of L11M0011. When measuring the agreement of two human annotators for fine granularity, the reliability level of all lectures are tentative.

4 Automatic Alignment between Slide Components and Lecture Utterances

Automatic alignment between slide components and lecture utterances will be required to realize automatic text summarization using slide components as conceptual units. This section explains our preliminary result of automatic alignment.

First of all, we formulate the automatic alignment problem between slide components and lecture utterances as the problem to find the mapping set M. A member of M is a single mapping m from a lecture utterance u to a slide component e ($u \rightarrow e$). Although there are many possible mapping sets, the eligible mapping set M must maximize the following objective function

$$f(M) = \lambda \sum_{m \in M} f_w(m) + (1 - \lambda) \sum_{m \in M} f_c(m, M), \tag{7}$$

where $f_w(m)$ represents the content-based agreement between the utterance u and the slide component e which are specified by the mapping m, and $f_c(m)$ represents the consistency score.

The content-based agreement score function $f_w(m)$ of the mapping m is defined as follows

$$f_w(m) = \frac{|N_u \cup N_e|}{|N_u \cap N_e|}, \tag{8}$$

Table 4: Result of Automatic Alignment (L11M0030)

λ	Accuracy		Recall		Overall accuracy
	I	O	I	O	
0	0.0896	0.656	0.113	0.734	0.247
0.25	0.329	0.693	0.424	0.734	0.425
0.5	0.335	0.693	0.432	0.734	0.429
0.75	0.341	0.697	0.440	0.734	0.434
1	0.248	0.699	0.321	0.728	0.365

where N_u is a set of nouns included in the utterance u specified by the mapping m, and N_e is a set of nouns included in the slide component e specified the mapping m. In this paper, the simplest agreement score function is employed as preliminary experiments, and it is future work to employ more sophisticated score function like (Guo and Diab, 2012).

Generally speaking, a common lecturer has a tendency to explain slide components in their appearance order. The latter member of the objective function $f(M)$ is designed to capture this tendency, and the consistency score function $f_c(m_i, M)$ is defined as follows:

$$f_c(m_i, M) = \begin{cases} -\sum_{j=0}^{i-1} \delta(e_i < e_j) & f_w(m_i) = 0 \\ 0 & otherwise \end{cases} \quad (9)$$

Suppose a mapping m_j which appears former than the certain mapping m_i in the utterance sequence. In other words, the utterance u_j specified by the mapping m_j precedes the utterance u_i specified by the mapping m_i. When the lecturer explains slide components in their appearance order, the slide component e_j specified by the mapping m_j precedes the slide component e_i specified by the mapping m_i consistently. The above function counts the number of mappings which do not meet this condition.

Table 4 shows the preliminary result of automatic alignment. λ is allowed to vary with the result of experiments, f_c has been found to not contribute significantly to the accuracy of the automatic alignment. Therefore, to further improve accuracy of the automatic alignment is needed improvements f_w.

5 Conclusion

This paper describes our developing corpus of lecture utterances aligned to slide components, which contains two contributions. The first contribution is to design the label system which represents alignment between textual units and slide components even when there are boundary mismatches between textual units and sentential boundaries. It is crucial inevitable problem to handle spontaneous speeches. The second contribution is to show the agreements between human annotators when the label system is employed. As a future work, we are going to investigate automatic decision of granularity level of slide components.

Acknowledgements

This work was supported by JSPS KAKENHI Grant No. 15K12097 and No. 25280062.

References

Yuya Akita, Masahiro Saikou, Hiroaki Nanjo, and Tatsuya Kawahara. 2006. Sentence boundary detection of spontaneous Japanese using statistical language model and support vector machines. In *Proceedings of INTERSPEECH*, pages 1033–1036.

M. Alley, M. Schreiber, and J. Muffo. 2005. Pilot testing of a new design for presentation slides to teach science and engineering. In *Frontiers in Education, 2005. FIE '05. Proceedings 35th Annual Conference*, pages S3G–7, Oct.

Jean Carletta. 1996. Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, 22(2):249–254, 6.

Timothy Chklovski and Rada Mihalcea. 2003. Exploiting agreement and disagreement of human annotators for word sense disambiguation. In *Proceedings of the Conference on Recent Advances in Natural Language Processing (RANLP2003)*.

Elena Filatova and Vasileios Hatzivassiloglou. 2004. A formal model for information selection in multi-sentence text extraction. In *Proceedings of Coling 2004*, pages 397–403, Geneva, Switzerland, Aug 23–Aug 27. COLING.

Jade Goldstein, Vibhu Mittal, Jaime Carbonell, and Mark Kantrowitz. 2000. Multi-document summarization by sentence extraction. In *Proceedings of the 2000 NAACL-ANLPWorkshop on Automatic summarization-Volume 4*, pages 40–48. Association for Computational Linguistics.

Weiwei Guo and Mona Diab. 2012. Modeling sentences in the latent space. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 864–872, Jeju Island, Korea, July. Association for Computational Linguistics.

Tessai Hayama, Hidetsugu Nanba, and Susumu Kunifuji, 2008. *PRICAI 2008: Trends in Artificial Intelligence: 10th Pacific Rim International Conference on Artificial Intelligence, Hanoi, V ietnam, December 15-19, 2008. Proceedings*, chapter Structure Extraction from Presentation Slide Information, pages 678–687. Springer Berlin Heidelberg, Berlin, Heidelberg.

Eduard Hovy, Chin-Yew Lin, Liang Zhou, and Junichi Fukumoto. 2006. Automated summarization evaluation with basic elements. In *Proceedings of the Fifth Conference on Language Resources and Evaluation (LREC 2006)*, pages 604–611.

N. Kitaoka, T. Yamada, S. Tsuge, C. Miyajima, T. Nishiura, M.Nakayama, Y. Denda, M. Fujimoto, K. Yamamoto, T. Takiguchi, S.Kuroiwa, K. Takeda, and S. Nakamura. 2006. CENSREC-1-C: Development of evaluation framework for voice activity detection under noisy environment. In *IPSJ technical report, Spoken Language Processing (SIG-SLP), Vol.2006, No.107*, pages 1–6.

Julian Kupiec, Jan Pedersen, and Francine Chen. 1995. A trainable document summarizer. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 68–73. ACM.

Hwee Tou Ng, Chung Yong Lim, and Shou King Foo. 1999. A case study on inter-annotator agreement for word sense disambiguation. In *Proceedings of the ACL SIGLEX Workshop on Standadizing Lexical Resource (SIGLEX99)*, pages 9–13.

N. Otsu. 1979. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man and Cybernetics*, SMC-9(1):62–66.

Dragomir R Radev, Hongyan Jing, and Malgorzata Budzikowska. 2000. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In *Proceedings of the 2000 NAACL-ANLP Workshop on Automatic summarization*, pages 21–30. Association for Computational Linguistics.

Kazuya Shitaoka, Kiyotaka Uchimoto, Tatsuya Kawahara, and Hitoshi Isahara. 2004. Dependency structure analysis and sentence boundary detection in spontaneous japanese. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hiroya Takamura and Manabu Okumura. 2009. Text summarization model based on maximum coverage problem and its variant. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 781–789. Association for Computational Linguistics.

Katsuya Takanashi, Takehiko Maruyama, Kiyotaka Uchimoto, and Hitoshi Isahara. 2003. Identification of "sentences" in spontaneous Japanese - detection and modification of clause boundaries. In *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, pages 183–186.

Masatoshi Tsuchiya, Satoru Kogure, Hiromitsu Nishizaki, Kengo Ohta, and Seiichi Nakagawa. 2008. Developing corpus of Japanese classroom lecture speech contents. In European Language Resources Association (ELRA), editor, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*.

Wen-tau Yih, Joshua Goodman, Lucy Vanderwende, and Hisami Suzuki. 2007. Multi-document summarization by maximizing informative content-words. In *IJCAI*, volume 7, pages 1776–1782.