Alignment between a Technical Paper and Presentation Sheets Using a Hidden Markov Model

Tessai Hayama*, Hidetsugu Nanba† and Susumu Kunifuji*
* Japan Advanced Institute of Science and Technology
1-1, Asahidai, Nomi-shi, Ishikawa-ken
Email: {t-hayama, kuni}@jaist.ac.jp
†Hiroshima City Univercity
3-4-1, Ozuka-higashi, Asaminami-ku, Hiroshima-shi, Hiroshima-ken
Email: nanba@its.hiroshima-cu.ac.jp

***Abstract*— We have been studying the automatic generation of presentation sheets from a technical paper. Our approach consists of obtaining a set of rules for generating presentation sheets by applying machine learning techniques to many pairs of technical papers and their presentation sheets collected from the World Wide Web. As a first step, in this paper, we propose a method for aligning technical papers and presentation sheets. Our method is based on Jing's method, which uses a Hidden Markov Model (HMM). Although her method is useful to align short sentences in newspaper articles, it is inapplicable to align sentences in a paper including charts and long sentences. Therefore, we analyse features of papers and sheets, such as information from text appearance, and propose an alignment method that combines the use of these features and her method. The evaluation shows that our alignment method performed effectively.**

I. INTRODUCTION

We have been studying the automatic generation of presentation sheets from a technical paper. Our approach consists of obtaining a set of rules for generating presentation sheets by applying machine learning techniques to numerous pairs of technical papers and their presentation sheets collected from the World Wide Web. As a first step, in this paper, we propose a method to align a technical paper and its presentation sheets.

Several methods to generate presentation sheets from a technical paper have been proposed. Shibata et al. focused on discourse structure of research papers, and devised several rules to generate presentation sheets [4]. Yasumura et al. proposed a method to extract important sentences, figures, tables and equations from a technical paper, and applied templates to them to generate presentation sheets [6]. These methods have been confirmed effective to some extent in their experiments. Towards fully automatic generation of presentation sheets, however, further improvement is required. It is quite time consuming to compare numerous pairs of papers and sheets manually and to look for useful patterns for generation of sheets from them. In this paper, we will therefore use machine learning techniques to obtain rules

In recent years, many research papers and their presentation sheets have become available on the World Wide Web. We have been constructing PRESRI[1], a research paper database, by collecting Postscript and PDF files from the Web [3]. Currently, PRESRI contains more than 90,000 papers written in Japanese or English. In addition to these files, we also have more than 10,000 PowerPoint files (presentation sheets) from the Web.

Before applying machine learning to these data, we must first detect pairs of technical papers and presentation sheets, then align sections in the papers and sheets automatically. In the first step, we detect pairs of papers and sheets by taking account of their locations on the Web and their titles. In the second step, we must consider the differences of expression between papers and sheets. Several studies have been related to this step. Uchimoto et al. proposed a method to align technical papers and their speech transcriptions [5]. They used n-grams with evaluation values and did not consider precise alignment, such as word-to-word and phrase-to-phrase. However, as detailed alignment is necessary for acquiring our automatic generation rules, we must use other alignment methods. Katoh et al. proposed a method to acquire automatic transformation rules from TV news texts and their teletexts for automatic generation of a teletext from a TV news text [2]. They used DP-matching to find the most likely sequences of characters between TV news texts and their teletexts. However, this method also is not applicable in our case, because presentation sheets are often ordered differently to the sections in technical papers.

One way to cope with this problem is to use a Hidden Markov Model (HMM). To obtain summarization rules, Jing proposed a method to align sentences in a newspaper article with sentences in its summary using a HMM [1]. In contrast to DP-matching, her approach aligns word orders in articles and their summaries well, even when the word orders differ. We therefore used a HMM for our study.

In the next section, we introduce Jing's method and point out a problem with her method when it is applied to our case. In Section Ⅲ, we propose a method to solve the problem described in Section Ⅱ. To investigate the effectiveness of our method, we conducted an experiment. We report our experiment and discuss the results in Section Ⅳ, and conclude our work in Section V.

[1] http://www.presri.com

II. Alignment using Jing's method

A. Jing's method

Jing's method aligns sentences in a summary with original-document sentences using a HMM. Her method determines the most likely position in the document for each word in the summary using a set of heuristic rules. Her method consists of the following steps.

First, the word sequence in a summary is formulated to make sets of states in the HMM as follows. A word's position in a document is uniquely represented by a sentence position and the word's position within the sentence: (Sentence Position, Word Position). For example, (2, 5) uniquely refers to the fifth word in the second sentence. Next, for each word in the summary, a corresponding position is located in the document. For example Fig. 1 indicates that a word "position" in the summary appears twice ((3, 2) and (21, 4)) in the document. In the summary, every time a different word position in each position sequence is chosen, a different summary sequence is obtained.

Second, the values of P1–P6 are assigned to transition probabilities between every possible pair of positions using the following heuristic rules.

- If $((SN_1 == SN_2) and (TN_1 = TN_2 - 1))$, then P1
- If $((SN_1 == SN_2) and (TN_1 < TN_2 - 1))$, then P2
- If $((SN_1 == SN_2) and (TN_1 > TN_2))$, then P3
- If $(SN_2 - CONST < SN_1 < SN_2)$, then P4
- If $(SN_2 < SN_1 < SN_2 + CONST)$, then P5
- If $(|SN_2 - SN_1| >= CONST)$, then P6

where SN_1 and SN_2 indicate two adjacent sentences in the document, TN_1 and TN_2 indicate two adjacent words within a sentence, $CONST$ is defined as a small constant , such as 3 or 5, and P1–P6 are experimentally assigned values. In Jing's experiment, P1–P6 were assigned evenly decreasing values: 1, 0.9, 0.8, and so on. Fig. 2 shows a graphical representation of the above rules for assigning bigram probabilities.

Finally, Jing's method uses the Viterbi algorithm to find the most likely document positions for each word in the summary sequences, which are shown as bold lines in Fig. 1.

B. Problems with Jing's method

Although Jing's method is useful for aligning sentences in a summary with those in a newspaper article, it is inapplicable in our case for the following three reasons.

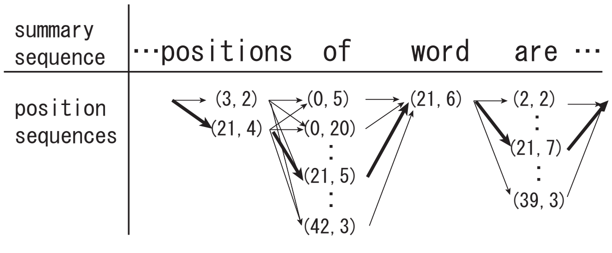

Fig. 1. The sequences of positions in summary sentences

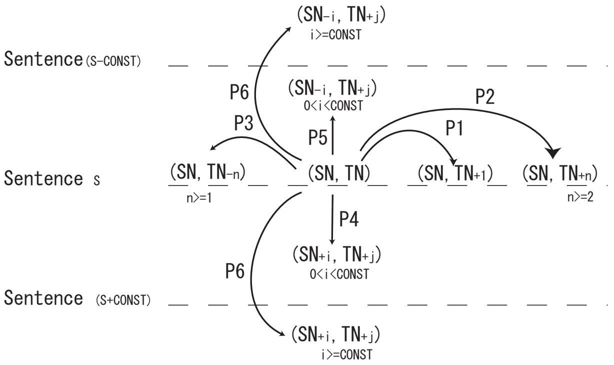

Fig. 2. Assigning transition probabilities in the HMM

Reason 1:
Jing's method attempts to align all words in a technical paper, and this causes alignment errors, because conjunctive expressions or function words such as case particles in a technical paper are often deleted or paraphrased to other expressions in presentation sheets. For examples, a Japanese conjunctive expression "Sokode (therefore)" is often replaced by a right arrow in a sheet. Some phrases that come at the end of a sentence are often shortened or deleted. For example, if a sentence ends with a sahen verb followed by its inflection, or helping verbs, it does not change the meaning much even if the verb stem (or sahen noun) is kept and deleted the rest of it.

Reason 2:
Insertion of additional sentences in presentation sheets also causes alignment errors. In presentation sheets, examples are often used instead of the abstract explanation in the technical paper.

Reason 3:
In contrast to newspaper articles, similar word sequences are used repeatedly in technical papers. That is, there are many candidates for alignment for one phrase in a sheet. In this case, it is difficult to determine the correct assignment using Jing's heuristic rules.

In the next section, we propose an alignment method that takes these problems into account.

III. Improvement of Jing's method

We improve Jing's method in the following four points: "Re-evaluation of alignment," "Length of phrase sequence," "Position gap," and "Alignment using titles." Here, the first three points are the improvements of reasons one, two and three, respectively, and the final one is the method using a feature of sheets and papers. In this section, we describe these points and show how to combine them with Jing's method.

A. Re-evaluation of alignment

We evaluate the degree of alignment between each section of the paper and the sheets, so we can assign the section to a sheet. To calculate the degree of alignment, it is effective to use representative words and title phrases in the sheet and the section. A sheet is likely to be assigned to a section when some words in the sheet appear only in the section, or when the phrase in the sheet title appears in sentences in the section.

Jing's method cannot consider such features of words. Therefore, we calculate the degree of alignment in each section by taking account of the following three points from the output of her method, as shown in Fig. 3. Firstly, we pay attention to representative words that appear only in particular sections. These words are useful to differentiate appropriate candidate sections from others. Secondly, we take account of sequence of words in sections. If a series of words in a presentation sheet also appear in the same sequence in a section, the section is probably the counterpart of the sheet. On the contrary, if a series of words in a sheet appear separately in a section, the section may not be the counterpart. Thirdly, we focus on the title words. A sheet and its corresponding section tend to contain a lot of common words in their titles. The equations for re-evaluation of alignment using these points are defined as follows:

$$Tv_{(word_{(i)})} = \left(\frac{Number\_of\_Sections(word_{(i)})}{Number\ of\ All\ Sections} \right)^2 \tag{1}$$

$$Lv_{(word_{(i)})} = \frac{1\ -\ Rel\_Value(word_{(i-1)}, word_{(i)})}{(Length\_of\_Phrase(word_{(i)}))^2} \tag{2}$$

$$Pv_{(word_{(i)})} = \begin{cases} p\_val & (if\ word_{(i)}\ in\ Title) \\ 1 & (if\ not\ word_{(i)}\ in\ Title) \end{cases} \tag{3}$$

$$ReEval = \sum_{i=0}^{tn-1} Tv_{(word_{(i)})} * Lv_{(word_{(i)})} * Pv_{(word_{(i)})} \tag{4}$$

where $word_{(i)}$ is defined as the ith word in the assigned word sequence and tn is the number of assigned word positions. $ReEval$ represents the degree of the alignment between the section and the sheet. $Tv_{(word_{(i)})}$, $Lv_{(word_{(i)})}$, and $Pv_{(word_{(i)})}$ represent the representativeness of $word_{(i)}$ for a section, the most likely sequence of words, and the importance of the position including $word_{(i)}$, respectively. $Number\_of\_Sections(word_{(i)})$, $Rel\_Value(word_{(i-1)}, word_{(i)})$, and $Length\_of\_Phrase(word_{(i)})$ are defined as the number of sections including $word_{(i)}$, the value following the same rules as Jing's transition probabilities, and the length of the phrase sequence including $word_{(i)}$.

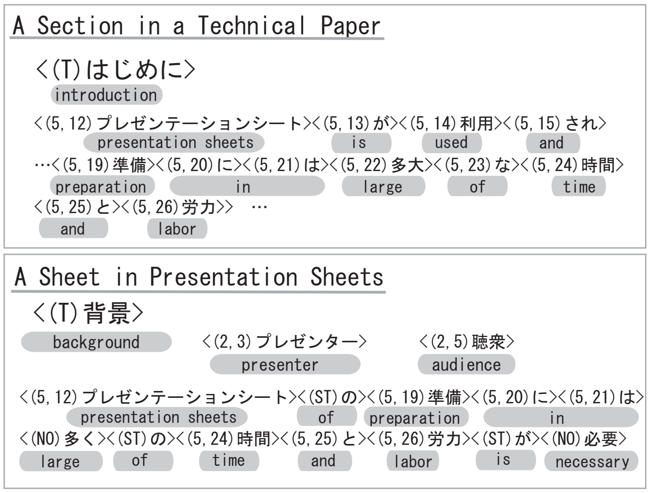
A Section in a Technical Paper
<(T)はじめに>
introduction
<(5,12)プレゼンテーションシート><(5,13)が><(5,14)利用><(5,15)され>
presentation sheets
is
used
and
…<(5,19)準備><(5,20)に><(5,21)は><(5,22)多大><(5,23)な><(5,24)時間>
preparation
in
large
of
time
<(5,25)と><(5,26)労力>> …
and
labor
A Sheet in Presentation Sheets
<(T)背景>
background
<(2,3)プレゼンター>
presenter
<(2,5)聴衆>
audience
<(5,12)プレゼンテーションシート><(ST)の><(5,19)準備><(5,20)に><(5,21)は>
presentation sheets
of
preparation
in
<(NO)多く><(ST)の><(5,24)時間><(5,25)と><(5,26)労力><(ST)が><(NO)必要>
large
of
time
and
labor
is
necessary

Fig. 3. Result of sheet alignment using Jing's method. Numbers in parentheses and the phrase in the brown rectangle represent position sequence in a paper and the meaning in English

An example of our method is shown in Fig. 4. In Fig. 4, there are five selected words; "Introduction" is a word in the sheet title, "propose" is a word that is not assigned any word positions in the section, and "personalized environment" is a phrase sequence including two word positions. $Number\ of\ ALL\ Sections$, p_val, and the starting value of $ReEval$ are set to 5, 0.5, and 1, respectively. The degree of alignment between the section and the sheet is given as 0.000002916. by multiplying the values of Tv, Pv, and Lv for each word. $ReEval$ provides a smaller value, as it aligns better between a section and a sheet; for example, if we calculate the score of ReEval between a sheet and its corresponding section, it will be nearly zero.

B. Length of phrase sequence

Jing's method assumed that all sentences or phrases in a summary are contained in the original document. As a result, the method makes errors whenever additional sentences are inserted in presentation sheets. In order to improve such errors, we modify the results of Jing's method by taking account of the length of phrase sequence. We assume that the longer the phrase sequence in a sheet is, the more appropiate the alignment is. We use the following rules: 1) In the sheet sequence, choose the word position included in the longest phrase sequence among them; 2) If there is more than one word position in the sheet sequence included in the longest phrase, choose the word position that is the nearest to the average word positions and is in the longest phrase among them. The matching is cancelled before applying these rules if an isolated function word and an initial function word in a phrase position are assigned.

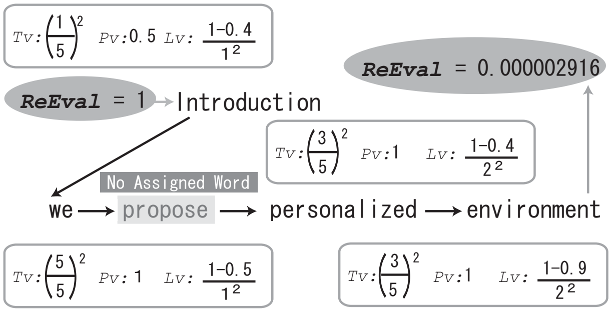
Tv: (1/5)^2 Pv:0.5 Lv: 1−0.4 / 1^2
ReEval = 0.000002916
ReEval = 1
Introduction
Tv: (3/5)^2 Pv:1 Lv: 1−0.4 / 2^2
No Assigned Word
we
propose
personalized
environment
Tv: (5/5)^2 Pv: 1 Lv: 1−0.5 / 1^2
Tv: (3/5)^2 Pv:1 Lv: 1−0.9 / 2^2

Fig. 4. Relating sections to sheets using the re-evaluation of alignment

C. Alignment using position gaps

To reduce the number of sentences in sections aligned with sentences in sheets, we choose some sections in the paper using each sheet position in the sheet sequence. Because researchers usually make presentation sheets according to the organization of the paper, the sequence of presentation sheets roughly corresponds to that of the sections in the paper. Therefore, we use the position gap, which represents the degree of positional difference between the sheet's position in the sheet sequence and the section's position in the section sequence. If the position gap is set appropriately, the corresponding section can be included within the range of the position gap based on the position sequence of each sheet.

The preparation of the alignment using the position gap involves normalizing each sheet position in the sheet sequence and section position in the section sequence within the range from 0 to 1. Every time we try to align sentences in the sheet with sentences in the sections, our method chooses some sections within the range based on normalized positions and the position gap range, as shown in Fig. 5.

D. Alignment using titles

Alignment using titles assigns each sheet to a section using phrases in the titles of both section and sheet.

The rules for alignment using titles are as follows:

R1 If titles of both a section and a sheet contain one of the following cue phrases, the sheet is assigned to the section.
- Set A. . .'Introduction', 'Background', and so on
- Set B. . .'Future Work(s)', 'Future Study(ies)', 'Future Research Direction', and so on
- Set C. . .'Conclusion', 'Summary', and so on

R2 If nothing is assigned to the second sheet in procedure R1, the sheet is assigned to the first section.

R3 If nothing is assigned to the sheet in procedure R1 and the title belongs to either Set B or C, the sheet is assigned to the final section.

R4 If nothing is assigned to the final sheet in procedure R1, the sheet is assigned to the final section.

The method above can be applied only when the sheet is the second or final position, or when the sheet includes a cue phrase in the title.

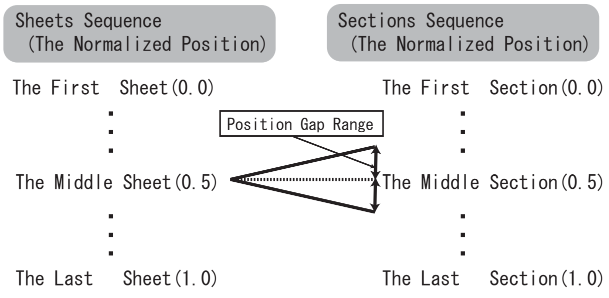

Fig. 5. Example of alignment using position gap

E. Combination of Jing's method and features of a paper and sheets

In this section, we show how we combine Jing's method with the four methods described above. The procedure consists of five stages as shown in Fig. 6. In the first stage, the most probable section for a sheet is sought from a paper using alignment using titles. If the rule finds a section for the sheet, the process finishes. In the second stage, some candidate sections for a sheet are selected from a paper using alignment using position gap. In the third stage, words in the sheet are assigned words in each section using Jing's method. In the fourth stage, the matching of inappropriate words is cancelled using length of phrase sequence. In the final stage, the probability of each sheet–section association is estimated by re-evaluation of alignment, and finally the sheet is assigned to the section that has the minimum degree of alignment.

IV. Experiment

A. Method and preparation

An experiment was performed to evaluate the effectiveness of our alignment method. We compared our method with Jing's original method according to the accuracy of the alignment. The stand-alone Jing's method in the experiment aligned sentences in each sheet with sentences in the paper using her method, and then assigned each sheet to the section including the words that correspond most.

We used 49 pairs of papers and presentation sheets, which were written in Japanese. The research field of the data used in our experiment is mainly in information science. The average number of pages in papers and sheets were 5.6 and 23.6, respectively. The paper data and the sheet data were captured automatically from Web materials, such as Postscript, PDF,

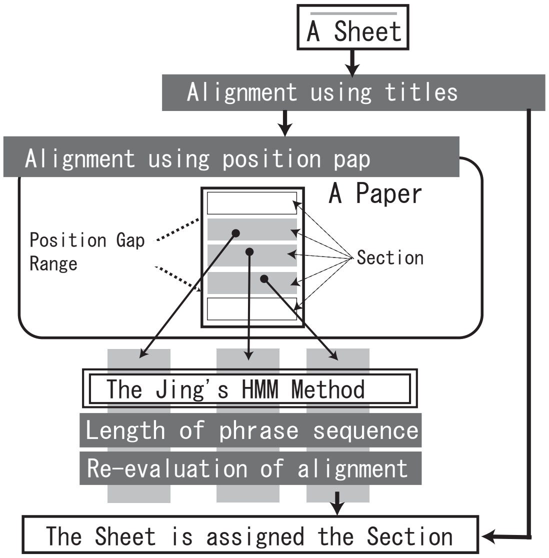
A Sheet
Alignment using titles
Alignment using position pap
A Paper
Position Gap Range
Section
The Jing's HMM Method
Length of phrase sequence
Re-evaluation of alignment
The Sheet is assigned the Section

Fig. 6. Processing flow of a combined alignment approach method using features of a paper and its sheets with Jing's HMM method

and PPT files[2]. The paper data was created as follows:

1) Convert Postscript files into PDF using ps2pdf[3].
2) Convert PDF files into XML using pdftohtml[4]. The XML files include information about character sizes and positions.
3) Find the borders of sections and the captions of figures and tables using the information in the XML files.
4) Insert these captions in the section explaining each figure and table.

The sheet data was created as follows:

1) Convert PPT files into plaintext with information about character sizes, positions, and sheet borders using a specially created filtering program.
2) Filter out sheets not used in the presentation, such as supplementary sheets, taking account of cue phrases and sheet positions.

We also created correct alignments manually. We investigated the relationship between the position gap range (PGR) and the number of sheets that could be aligned with sections within the gap. The results in Table I show that sheets and sections could be aligned in 98% of cases when the parameter is set to 0.5. ASR_in_HJD is 1%, when PGR is 0.9. The most of these cases are the positional differences of related researches between papers and sheets. But as they are rare cases, we ignore them. We therefore set the PGR to be less than 0.5.

B. Results and discussion

We evaluated our method using two kinds of data: ALL and CDI. ALL evaluates our method using all the data, while CDI is part of ALL. The ratio of the number of sheets in CDI among ALL is 91% (1011/1133). As we described in Section II-B, some presentation sheets cannot be aligned to any sections in a paper. We therefore eliminate these sheets from ALL (CDI), and evaluate our method using CDI.

In our experiments, Jing's original method obtained accuracies of 63.2 and 67.2% in ALL and CDI, respectively. The results for our method are shown in Table II. The accuracy of our method was higher than Jing's method. When the position gap is 0.30, we obtained the best scores.

Most of our errors were caused by a small quantity of textual information in the sheets. In alignment results for such sheets, there were few correspondence relationships to papers. As our final purpose is to acquire rules for automatically generating sheets from a paper using the alignment results, it would be necessary to specify and exclude such sheets from the results based on the re-evaluation of alignment.

TABLE I
RELATIONSHIP BETWEEN POSITION GAP RANGE AND FRACTION OF SHEETS THAT COULD BE ALIGNED

PGR	ASR_in_HJD	PGR	ASR_in_HJD
0.0	55%	0.5	1%
0.1	26%	0.6	0%
0.2	11%	0.7	0%
0.3	5%	0.8	0%
0.4	1%	0.9	1%

PGR: Position Gap Range,
ASR_in_HJD: fraction of sheets that could be aligned

TABLE II
EVALUATION OF THE ALIGNMENT METHOD COMBINED WITH THE METHOD USING FEATURES OF A PAPER AND SHEETS AND JING'S METHOD

PGR	Accuracy		PGR	Accuracy	
	ALL	CDI		ALL	CDI
no	76.1%	79.9%	0.35	79.3%	82.9%
0.20	77.5%	80.2%	0.36	79.2%	82.7%
0.21	78.5%	81.3%	0.37	79.2%	82.8%
0.22	78.2%	81.0%	0.38	79.2%	82.8%
0.23	78.4%	81.3%	0.39	79.2%	82.9%
0.24	78.6%	81.6%	0.40	79.0%	82.8%
0.25	79.1%	82.2%	0.41	78.9%	82.8%
0.26	79.2%	82.4%	0.42	78.7%	82.6%
0.27	79.3%	82.6%	0.43	78.7%	82.6%
0.28	79.6%	82.9%	0.44	78.7%	82.6%
0.29	79.5%	82.9%	0.45	78.5%	82.4%
0.30	**79.6%**	**83.0%**	0.46	78.3%	82.1%
0.31	79.6%	82.8%	0.47	78.2%	82.0%
0.32	79.5%	82.9%	0.48	78.2%	82.0%
0.33	79.4%	82.8%	0.49	77.8%	81.6%
0.34	79.5%	82.9%	0.50	77.8%	81.5%

ALL: All correct data, CDI: Correct Data that can be Identified

V. CONCLUSIONS

In this paper, we proposed a method to align presentation sheets with technical papers in Japanese. We improved Jing's method by taking account of features of papers and sheets. The experimental results showed that our method could improve Jing's method. In our future work, we will apply our method to numerous pairs of papers and presentation sheets, and obtain rules for generating presentation sheets using machine learning techniques.

REFERENCES

[1] H. Jing. Using hidden markov modeling to decompose human-written summaries. *Computational Linguistics*, 4(28):527–544, 2002.

[2] N. Katoh and N. Uratani. A new approach to acquiring linguistic knowledge for locally summarizing japanese news sentences (in Japanese). *Journal of Natural Language Processing*, 7(6):73–92, 1999.

[3] H. Nanba, T. Abekawa, M. Okumura, and S. Saito. Bilingual presri: Integration of multiple research paper databases. *In Proceedings of RIAO 2004*, pp. 195–211, 2004.

[4] T. Shibata, D. Kawahara, and S. Kurohashi. Syudai to Bunsyokozo no Kaiseki ni Motozuku Suraido no Zidoseisei (in Japanese). *In Proceedings of the 9th Annual Meeting of the Association for Natural Language Processing (NLP2003)*, pp. 597–600, 2003.

[5] K. Uchimoto, C. Nobata, K. Ohta, Q. Ma, M. Murata, and H. Isahara. Yokou to Sono Koenkakiokoshi no Taiouzuke oyobi Kakiokoshi no Tekisutobunkatsu (in Japanese). *In Proceedings of the 7th Annual Meeting of the Association for Natural Language Processing (NLP2001)*, pp. 317–320, 2001.

[6] Y. Yasumura, M. Takeichi, and K. Nitta. A support system for making presentation slides (in Japanese). *Transactions of the Japanese Society for Artificial Intelligence*, 18(4):212–220, 2003.

[2] file format of the PowerPoint tool of the Microsoft Co.

[3] http://www.cs.wisc.edu/~ghost/

[4] http://pdftohtml.sourceforge.net/