MULTIFORM SUMMARIZATION FROM PRODUCT SPECIFICATIONS

Kazutaka SHIMADA[*1] , Tetsuro ITO[*2] and Tsutomu ENDO[*1]

[*1] *Department of Artificial Intelligence, Kyushu Institute of Technology, Iizuka, Fukuoka 820-8502 Japan*
{*shimada, endo*}*@pluto.ai.kyutech.ac.jp*

[*2] *Department of Computer Science and Intelligent Systems, Oita University, 700 Dannoharu Oita, 870-1192 Japan*
ito@csis.oita-u.ac.jp

Specifications contain many data. It is not, however, clear which ones are the characteristic-data among them. We have reported a product ranking system using product specifications. The system extracts characteristic-data from specifications about computer systems. In this paper, we propose a method for generating sentences from the extracted characteristic-data and integrating the generated sentences and the specifications. We define sentence generation frames (SGF) and explanation sentences (ES) for sentence generation. Sentences are generated from the SGF of which the slots are filled with characteristic-data. The system integrates the generated sentences and the tables relevant to them. Experimental results show the effectiveness of our system.

Key words: Sentence Generation, Table, Summarization, Integration

1. INTRODUCTION

As the World Wide Web rapidly grows, a huge number of online documents are easily accessible on the Web. Finding information relevant to user needs has become increasingly important. One of the useful online documents is specifications for equipment about products such as personal computers and digital still cameras. In general, their specifications are presented in tabular form as shown in Fig. 1.

We are developing a multi-specifications summarization system from multiple Web sites, focusing on personal computer products [6]. The system analyzes product specifications presented in tabular form and extracts characteristic-data from the analyzed data using user's requests. Figure 2 shows the process flow of our system.

In this paper, we report a method for generating sentences from the extracted characteristic-data. A text is suitable to grasp a compendium of data on each product, but not suitable to express details of difference about each product. On the other hand, a table is suitable for comparing detailed information. Our system can present the specifications about all data of a product. Since comparing them requires a huge amount of work for users, specifications should be summarized for output. Considering legibility for users, generated sentences and summarized specifications should be integrated. We describe a method for the integration and evaluate the effectiveness.

2. RELATED WORK

There are several approaches to extract information using document structure such as itemization and tabular forms. Sato et al. have proposed a method for automatic generation of digests from the NetNews [5]. Kawai et al. have proposed a method for automatic extraction of relational information from itemized text [4]. However, they are different from table forms that we handle. Although Chen et al. have reported a method for mining tables from HTML documents, their systems

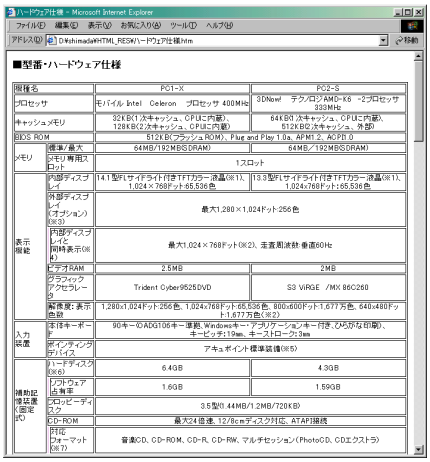


FIGURE 1. Specifications of products.

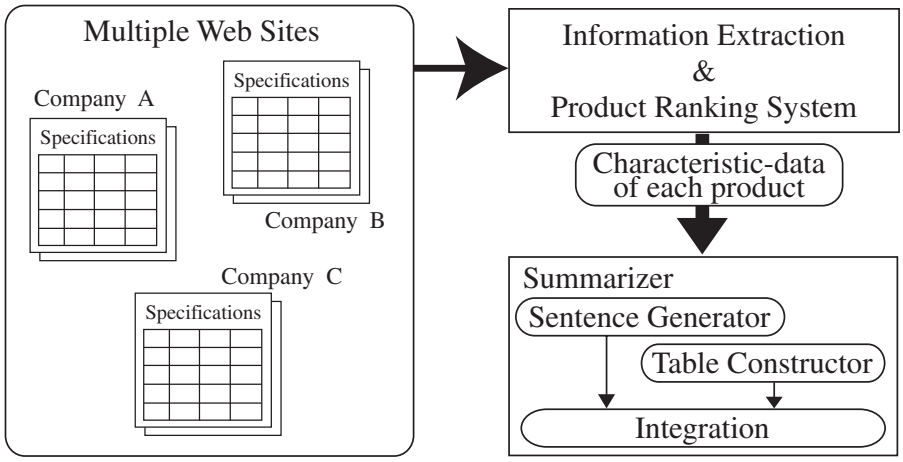


FIGURE 2. Outline of our system.

analyze only one table [2]. Wang et al. have reported a machine learning based approach for table extraction [9]. The purpose of Chen et al. and Wang et al. is to extract tables from the Web. They do not, however, deal adequately with the usage of structured data from tables. There is an approach to integrate several tables [11]. The purpose is to build ontologies from the World Wide Web via HTML tables.

There are many shopbots on the World Wide Web [3]. Most of them compare only the prices of products with each other. Chai et al. have proposed a conversational dialog system for online shopping [1]. The system deals with information of one site only. Also, the output of the system is a simple explanation sentence only. White et al. have reported a multidocument summarization system via information extraction [10]. The system outputs a text summary only. Our system can present a summary integrated texts and tables.

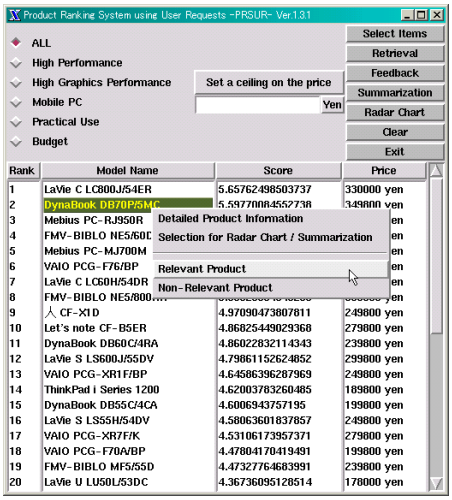


FIGURE 3. A prototype system.

3. PRODUCT RANKING SYSTEM

Although they contain many kinds of data, it is not clear which ones are the characteristic-data among them. For example, consider users who want to buy a personal computer. They retrieve product information that includes specifications from Web sites of many computer makers. However, it is difficult for users except some experts to select a suitable computer for their own purpose from the several specifications.

The purpose of our study is to develop a multimedia summarization system. As the initial step, we focus on a table on the World Wide Web. We are developing a multi-specifications summarization system from multiple Web sites [6]. Figure 3 shows a snapshot of our system. Our system has two processes: an information extraction (IE) and a product ranking process (see Fig. 2).

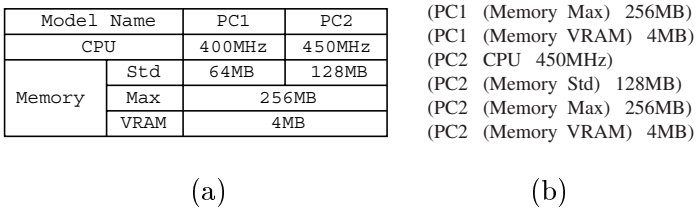
The IE process consists of table detection and table conversion. For table detection, our system extracts product specifications using weighted keywords. The weights of keywords are calculated by Bayes theorem. See [7] for details of the weighting and the table detection process. Next, we convert the specifications into table structures. A table structure is a set of simple ternary lists:

(Nam Atr Val)

Figure 4 (b) shows an example of table structures converted from HTML data (Fig. 4 (a)). An algorithm for the conversion is as follows:

1. Decompose a unified cell by HTML tags, ROWSPAN and COLSPAN (e.g., "Memory" in Fig. 4 (a)).
2. Convert each cell into table structure.
3. For each table structure, normalize Atr and Val. Ex. Monitor, Screen ⇒ Display

See [6] for details of table conversion.



Model Name		PC1	PC2
CPU		400MHz	450MHz
Memory	Std	64MB	128MB
	Max	256MB	
	VRAM	4MB	

(a)

(PC1 CPU 400MHz)
(PC1 (Memory Std) 64MB)
(PC1 (Memory Max) 256MB)
(PC1 (Memory VRAM) 4MB)
(PC2 CPU 450MHz)
(PC2 (Memory Std) 128MB)
(PC2 (Memory Max) 256MB)
(PC2 (Memory VRAM) 4MB)

(b)

Figure 4. An example of specifications and the table structures.

Table 1. Classification of Attributes.

Quantitative	Qualitative
CPU Clock: MHz, GHz	CPU Processor
Memory: MB	Graphics
Display: inch	CD-R/RW
Weight: kg	DVD-ROM, DVD-RAM
Dimensions: mm, cm	Pre-installed-OS
...	...

For the product ranking process, Our system extracts the attributes and values that characterize each PC. The attribute is classified into two categories: quantitative and qualitative. The typical example is listed in Table 1.

For quantitative attributes, the characteristic-data are extracted by comparing each value. The mean or mode of each attribute is computed from sample data. We call it a standard value. Each value obtains a score by comparing it with the standard value. We define the scores: minimum, standard, and maximum points are 0, 5, and 10 points respectively. Our system calculates the value per 1 point from them. The calculation is exemplified in Fig. 5. Assume that "500MHz", "600MHz", and "1.1GHz" are the minimum, standard, and maximum value, which were calculated from all PCs, respectively. If the clock speed of a PC is "800MHz", the PC obtains 7 points.

For qualitative attributes, we employ domain knowledge, which consists of keywords such as processor names, for the characteristic-data extraction. The number of keywords in domain knowledge is 23. The keywords possess weight.

To find PCs that relate to a user's request, we define relationships between a user's request and attributes. Table 2 shows examples of the relationships. Each attribute possesses weight. The score calculation process is as follows:

1. Select the table structures with attributes relating to a user's request.
2. For each selected PC, compute

$$score(c, r) = \frac{\sum_{k=1}^{n} w(a_k, r) \times pt(a_k, c)}{\sum_{k=1}^{n} w(a_k, r)} ,$$

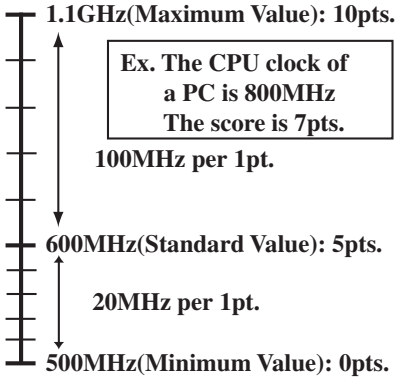


FIGURE 5. The calculation of the score.

TABLE 2. The relationships between a user's request and attributes

Request	Attributes
High performance	CPU, Memory, Display, HDD, Interface
High graphics performance	Display, Graphics, CPU, Memory
Mobile PC	Battery life, Dimensions, Weight
Practical use	CPU, HDD, Price, Interface, Software
Budget PC	Price, Software, Memory, CPU

where c, r and a_k are a PC, a user's request and an attribute respectively. $w(a_k, r)$ is the weight of a_k in the request. We define $w(a_k, r)$ manually. $pt(a_k, c)$ is the score calculated in previous paragraphs (see Fig. 5).

3. Return them as the recommended computers in descending order for the score.

4. GENERATION OF JAPANESE SENTENCES AND SUMMARIZED TABLES FROM SPECIFICATIONS

Our system can generate Japanese sentences and summarized tables from the characteristic-data of selected products by a user. We define sentence generation frames (SGF) and explanation sentences (ES) for sentence generation. Figure 6 shows document structure in our system. The number of topics is 9. Table 3 shows relationships between the topics and attributes in specifications. Examples of the SGFs are as follows:

- Topic SGF:
  [Topic] no {sugureta or yoi} [Nam]. (Japanese)
  [Nam] is excellent in [Topic]. (English)

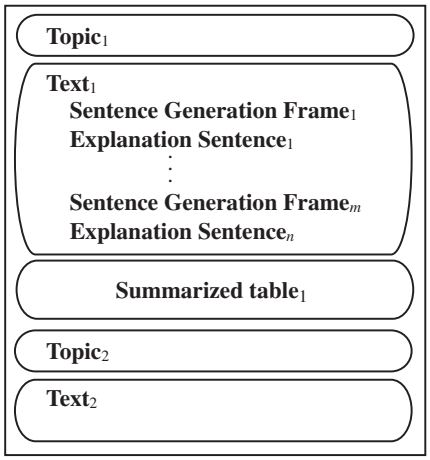


Figure 6. Document structure.

Table 3. Relationships between the topics and the attributes.

Topic	Attributes
Performance	CPU, Memory, Hard disk, etc.
Scalability	PCI, USB, PC card, etc.
Image processing	Graphics Chipset, Image processing soft, etc.
Display	Screen size, Resolution, VRAM, etc.
User-friendliness	Key size, Input device, etc.
Mobility	Weight, Dimensions, Battery life, etc.
Communication	Modem type, LAN, etc.
Sound	Speaker, Sound board, etc.
Soft	OS, Bundled software, etc.

- Text SGF:
  [Nam] wa [Atr] ni [Val] wo {tousai or saiyou}. (Japanese)
  [Nam] is equipped with [Atr] of [Val]. (English)

[Topic], [Nam], [Atr], and [Val] are slots. Sentences are generated from the SGF of which the slots are filled with characteristic-data. The number of SGFs is 27 frames. The slots possess constraints. The constraints are as follows:

- AllA : A slot can be filled with any attributes.
- AllV : A slot can be filled with any values.
- SA : A slot can be filled with specific attributes.
- SV : A slot can be filled with specific values.

An example is as follows:

[Nam] wa [SA-:USB|Serial|IEEE] port wo [SV-:[0-9]+] ko soubisiteiru. (Japanese)
The number of [SA-:USB|Serial|IEEE] ports on the [Nam] is [SV-:[0-9]+]. (English)

The SGF denotes that the 2nd slot (the 1st slot in the English example) is filled with USB, Serial or IEEE and the 3rd slot is filled with numerals.

ESs are employed to generate additional information to supplement for the generated sentence by SGFs. ESs possess the condition for generation, but do not possess any slots. The number of ESs is 35 sentences. An example of the ESs is as follows:

- Condition: [Atr] = "USB"
- ES: USB ha syuhenkiki wo tunagu interface desu. (Japanese)
  Universal Serial Bus, or USB, is an interface for connecting peripherals to your PC. (English)

We call the [Nam] of a Topic SGF a topic model (TM). It is determined as follows:

$$TM = argmax \sum_{a_i \in Topic} Score(C_j, a_i) \ ,$$

where a_i and C_j are an attribute and a product respectively. $Score(C_j, a_i)$ is the score of an attribute.

The sentence generation process using SGFs and ESs is as follows:

1. Extract table structures including characteristic data.
   Here characteristic-data is that the score is 10.
2. Classify the table structures by topics.
3. Classify the table structures by products.
4. Sort the table structures, based on the order in specifications.
5. Output the sentences about the TM preferentially.
6. Output the ES if the condition permits.

If the score calculated in the product ranking process is more than 5, our system generates an additional sentence such as "[Nam] ha hikakuteki yoi ([Nam] is comparatively good)." Figure 7 shows an example of the sentence generation process.

Next, the system refines the generated sentences using rules. The number of rules is 8. An example of the rule is as follows:

- Integration:
  Condition: [Nam$_i$] != [Nam$_{i+1}$] & [Atr$_i$] = [Atr$_{i+1}$]
  Process: The two sentences are integrated.

Figure 8 shows an example of the refinement process.

A text is not suitable for expressing all information. On the other hand, a table is suitable for comparing detailed information. However, presentation of real specifications is not appropriate because of legibility for users. Our system restructures summarized specifications for each Topic and integrates them and generated sentences. The table reconstruction process is as follows:

1. Extract table structures according to topics.

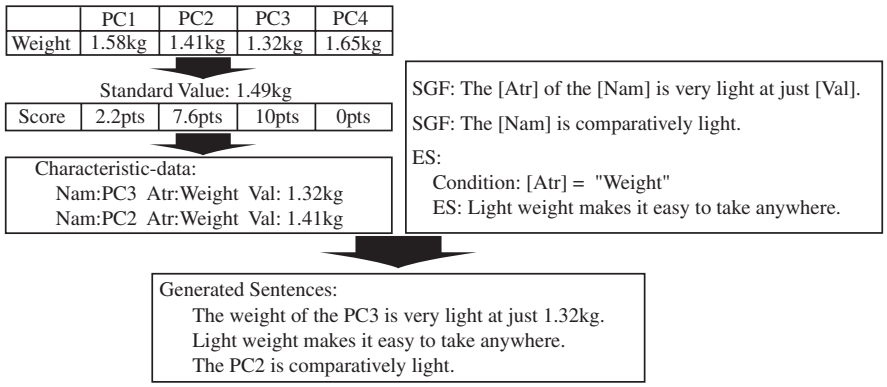


Figure 7. Sentence generation processing.

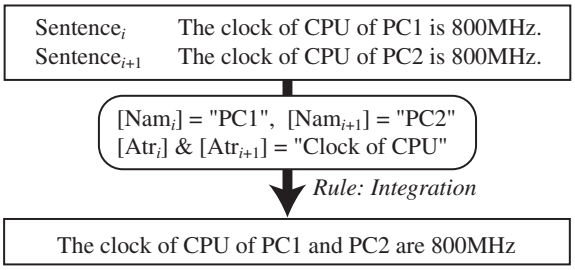


Figure 8. Refinement of generated sentences.

2. Unify the table structures.
3. Allocate tags such as "colspan" and "rowspan" in HTML to each attribute and each value where necessary.

5. EVALUATION

First, we evaluated significance of multiform summarization with 8 graduate students. The 8 graduate students evaluated the following 4 output forms:

Form (1) Specifications including characteristic-data only.
Form (2) Specifications including highlighted characteristic-data.
Form (3) Generated Sentences from characteristic-data.
Form (4) Generated sentences from characteristic-data, and summarized tables.

TABLE 4. Significance of multiform summarization.

Form (1)	Form (2)	Form (3)	Form (4)
10 pts.	23 pts.	18 pts.	29 pts.

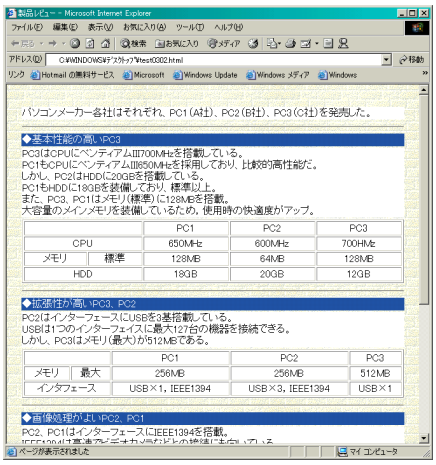


FIGURE 9. Generated sentences and tables.

We allocated 4 points for the best output form, and so 2nd is 3 points, 3rd is 2 points and 4th is 1 point. Table 4 shows experimental results. They evaluated the **Form (4)** as the most readable information. The experimental result shows the significance of integration of information.

Next, we evaluated generated sentences with 8 graduate students. Figure 9 shows an example of the generated sentences. We employed 3 generated documents for the evaluation. The evaluation criteria of the generated sentences were as follows:

Eval (1) The grammatical accuracy.
Eval (2) The usage of conjunction.
Eval (3) The textual coherence.
Eval (4) The redundancy of expression.
Eval (5) The legibility.

The 8 graduate students classified the generated sentences into the following 5 categories:

Bad : 1 point.
Below average : 2 points.
Fair : 3 points.
Good : 4 points.
Excellent : 5 points.

Table 5. Evaluation of generated sentences.

Eval (1)	Eval (2)	Eval (3)	Eval (4)	Eval (5)
3.7	2.9	3.7	4.3	4.0

Table 5 shows experimental results. There was room for improvement, especially **Eval (2)**. It was caused by rules for refinement of sentences. The system needs the improvement and addition of the rules. However, generated sentences work as summarization.

6. CONCLUSIONS

In this paper, we proposed a method for generating sentences from the extracted characteristic-data and integrating the generated sentences and the specifications. The proposed system presents readable information as a summary to users.

Future work will include (1) construction of domain knowledge by machine learning, (2) information retrieval through man-machine dialogue, and (3) integration with other sources such as product images [8].

REFERENCES

[1] J. Chai et al., "The role of a natural language conversational interface in online sales: a case study," International Journal of Speech Technology, pp. 285-295, 2001.

[2] H.H. Chen, S.C. Tsai, J.H. Tsai, "Mining tables from large scale HTML texts," Proceedings of COLING2000, pp.166–172, 2000.

[3] R.B. Doorenbos, O. Etzioni, and D.S. Weld, "A scalable comparison-shopping agent for the World Wide Web," Proceedings of the first International Conference on Autonomous Agents, 1997.

[4] A. Kawai, T. Tsukamoto, K.Yamamoto, and T. Shiino, "Automatic extraction of relational information using document structure from itemization and tabular forms," IEICE Transactions Information and Systems, vol.J81-DII, no.7, pp. 1609–1620, 1998 (in Japanese).

[5] M. Sato, S. Sato, and Y. Shinoda, "Automatic digesting of the NetNews," Trans. IPSJ, Vol.36, No.10, pp. 2371–2379, 1995 (in Japanese).

[6] K. Shimada, A. Fukumoto, and T. Endo, "Information extraction from personal computer specifications on the Web using a user's request," IEICE Transactions on Information and Systems, vol.E86-D, no.8, Aug. 2003.

[7] K. Shimada, K. Hayashi, and Tsutomu Endo, "Keywords and weighting for product specifications extraction," Proceedings of PACLING2003, 2003.

[8] K. Shimada, T. Ito, and T. Endo, "Classification of Images using Their Neighboring Sentences," Proceedings of PACLING2001, pp. 250-256, 2001.

[9] Y. Wang and J. Hu, "A machine learning based approach for table detection on the Web," Proceedings of The Eleventh International World Web Conference, 2002.

[10] M. White et al., "Multidocument summarization via information extraction," Proceedings of HLT2001, 2001.

[11] M. Yoshida, K. Torisawa, and J. Tsujii, "Extracting ontologies from World Wide Web via HTML tables," Proceedings of PACLING 2001, pp. 332–341, 2001.