IJCAT - International Journal of Computing and Technology, Volume 1, Issue 11, December 2014
ISSN : 2348 - 6090
www.IJCAT.org

Research on Document Summary Generation Using Attribute Information

[1] Abdunabi Ubul, [2] Hidekazu Kakei, [3] Jun-ichi Aoe

[1] Faculty of Integrated Arts and Sciences, Tokushima University,
Minami josanjima 1-1. 770-8502 Tokushima, Japan

[2] Institute of Socio-Arts and Sciences, Tokushima University,
Minami josanjima 1-1. 770-8502 Tokushima, Japan

[3] Department of Information Science and Intelligent Systems, Tokushima University,
Minami josanjima 2-1 .770-8506 Tokushima, Japan

Abstract - Recently, the expansion of the Internet has led to a deluge of information on the Web, making it difficult for users to locate efficiently needed information. To facilitate efficient searching for information, research into technology that can summarize the general outline of a text document is essential. This is especially true on the Web, where information from bulletin boards, blogs, and other sources is being used as consumer generated media data. Hence, summarizing technology that can accurately capture opinions, impressions, and fields of discussion is necessary. However, research efforts thus far have yet to yield satisfactory results. In this paper, we propose a method for generating a summary document using three types of attribute information acquired from the original document: the field, associated terms, and by using attribute grammars that combine these three attributes in document generation, we establish a formal and efficient generation technology. Experiments using information from 400 blogs found that when including the field and sensibility attributes, the summary accuracy rate, readability, and meaning integrity are 88.7%, 85%, and 86%, respectively. In comparison with traditional technologies, these three evaluation criteria are each 4% higher, thus demonstrating the effectiveness of this method.

Keywords **- Blog Document, Field Association, Attributes Grammar, Sensibility.**

1. Introduction

In recent years, the rapid advance of Internet technology has led to a deluge of information on the Web, making it difficult to locate required information efficiently. Important sentence extraction method 7 that extracts a sentence important in the text and assumes the summary is given to nine as the main technique to summarize the document. [1, 2, 3, 4, 5, 6]

Automatic generic document summarization based on un supervised schemes is very useful approaches because of no training data. Zha[7]and Yeh[8] have proposed summarization methods using Latent Semantic Analysis (LSA), but the LSA methods cannot extract meaningful sentences because of many features with positive and negative values. Li [9] has solved this problem by introducing generic multi-document summarization has proposed the mutual reinforcement principle (MRP) to query-based document summarization. However, there remain problems to extraction of subtopics in summarization. In order to solve this problem, a new unsupervised generic document summarization method using a non-negative matrix factorization (NMF) method has been proposed [10] [11]. The semantic feature vectors extracted from NMF can be interpreted more intuitively than those extracted from LSA-related methods.

This is that is the technique assumed to be a summary by sequentially choosing a sentence important in the document until becoming desired length, and permuting in order of appearance in the document and outputting the extracted sentence . From which various methods are proposed as the occurrence rate and the location information of the word are used for the judgment of an important sentence. Because information on the original is used as it is in these techniques, the summary generation that introduces writer's impression and expression of feeling cannot be achieved. To solve this problem, it proposes the summary generation technique for excerpting an important sentence from the blog document by using attribute information on the field

association word and the sensibility expression, etc. used with the blog document in this thesis.

Especially, it aims achieving the speed-up of the summary generation in the attribute grammar and the reduction of the amount of the memory by applying to the generation technique of the attribute grammar in the summary generation. In Section 2, we present the method for acquiring the fields and Sensibility from the document to generate the summarized document. In Section 3, we describe the inclusion of attribute grammar to generate the summarized document. In Section 4, we present the results of experiments evaluating this method. In Section 5, we summarize the paper and discuss future tasks.

2. Overview of Summarizing Technology

Depending on the planned use of the generated summary, it could be used in place of the original document, or alternatively the summary could be used to determine whether or not the original document should be read: these different categories are known as informative summaries and indicative summaries, respectively. Depending on the method used to create the summary, key points in the original document may be exported and used in a word-for-word summary (i.e., an extract), or instead those points may be summarized in a modified form (i.e., an abstract). A user-focused summary is centered on the user, and the summary is created based on the user's interests. More general summaries are known as generic summaries, and no assumptions about the user are made when such summaries are created. This paper focuses on the creation of user-focused summaries for readers of informative summaries, abstracts, and blogs.

2.1 Summary Flow

Figure 1 shows the flow for the summary system that we propose in this paper. The fields are determined from the input document. Then, the modules for extracting the field association words that were used in the field determination, as well as the Sensibility of the input document, is determined. The modules that extract the Sensibility expressions used to determine the Sensibility are processed in parallel. Then, those results are used to generate the summary document based on the attribute grammar.

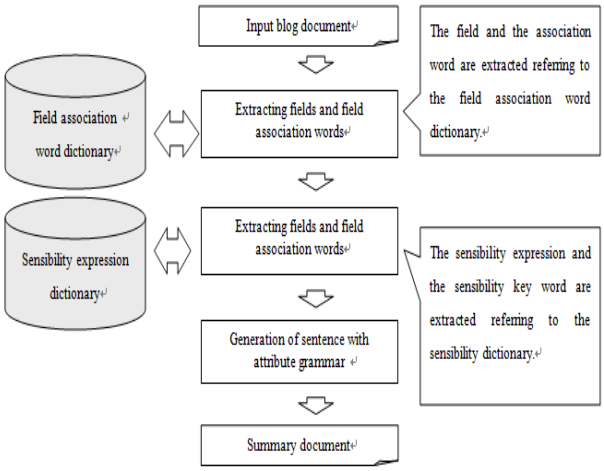

Fig. 1 Overall flow of summary system

2.2 Method for Determining Field Association Words

Words for which a particular field can be associated are called field association words [12] [13]. The association words are categorized according to is the strength of the association with the field. Table 1 presents the field association words and the corresponding field for each level (¥ indicates a high level field and low level field). We define the standard LEVEL as follows.

Definition 1. LEVEL (*w*) is the level of the field association words.

Level 1 [Complete association word]: A field association word w that can be uniquely associate with a terminal field is defined as LEVEL (*w*) = 1. For example, "Daisuke Matsuzaka" in Table 1 can be uniquely associated with the terminal field <baseball>, and thus is a complete association word.

Level 2 [Semi-complete association word]: A field association word w that can be associated with multiple terminal fields that belong to the same parent field. These are defined as LEVEL (*w*) = 2. For example, "racket" in Table 1 can be associated with multiple terminal fields including <tennis>, <table tennis>, and <badminton>.

Level 3 [Intermediate association word]: A field association words w that can be associated with only an intermediate field is defined as LEVEL (*w*) = 3. For example, "athlete" in Table 1 cannot be associated with any terminal fields; however, "athlete" can be uniquely associated with the intermediate field <sport> and thus is an intermediate association word.

Level 4 [Possible association word]: A field association words w that can be associated with multiple intermediate fields or terminal fields is defined as LEVEL (*w*) = 4. For example, "win or loss" in Table 1 can be associated with the intermediate field <sport> and the terminal field <shogi>, and thus is a possible association word.

Level 5 [Non-association word]: A non-association word cannot be associated with any field and is defined as LEVEL (*w*) = 5. For example, "case" in Table 1 cannot be associated with any particular field, and thus is considered a non-association word.

Table 1: Examples of field association words and levels

Field association word, *w*	Associated field, FIELD(*w*)	LEVEL(*w*)
"Daisuke Matsuzaka"	<Sport ¥Baseball>	1
"Racket"	<Sport ¥Tennis, table tennis, badminton>	2
"Athlete"	<Sport>	3
"Win or loss"	<Sport>, <Game ¥Shogi>	4
"Case"	None	5

2.2.1. Using Field Association Words to Identify Fields of a Blog

Here, let us provide a specific example to explain the use of field association words to identify the fields of a blog. The field identification flow is as follows. POINT (*f,w*) is the point value of association word w for field f, which is the set of field association words w for F-WORDS(*d*) of document d. The frequency of appearance, known as FREQ (*d,w*), is the frequency that association word w appears in document d. WEIGHT(*d,f*) is the weight of field f with respect to document d.

Algorithm 1. To determine fields f and the priority field of blog d.
Step 1: Determine the set of association words F-WORDS (d) for document d.
Step 2: Calculate POINT (*f,w*)*FREQ(d,w) for all elements of F-WORDS(*d*) which are association words w, and for all the elements of FIELD(*w*), which is field *f*. The summation of that calculated value is determined to be the weight f_WEIGHT (*d,f*) for field *f* with respect to document *d*.
Step 3: The field *f*, which had the highest summation of the calculated point total from the field determination Step 2, is determined to be the blog's priority field.

The following flow is demonstrated for blog *d* in Fig. 2, and the results are listed in Table 2, which shows the

I cannot buy the **year-end grand prize** lottery ticket because I have not reached the minimum age. It's interesting that they defeated the tough Kyojin team isn't it. Sometimes people **blog** using really surprising words, don't they? In 2007, the **Dragons** had a fierce match with Kyojin and Hanshin to take 2nd place in the **Central League**. But with the CS that was introduced last year, the Tigers and Giants exploded into the **Japan Series**. Then, in the **Japan Series** Nippon Ham got their revenge by taking a splendid first place victory with 4 wins and 1 loss. Can the pace be continued? Will there be no negative impact on Morino? Byon showed us some great batting against a thin fielding defense in the **exhibition match**. I wasn't expecting **Masa Yamamoto** for the **pitching staff**.
I didn't think it would end with them having just 2 wins. I thought they would get at least get 7 wins. I think that Hoshi who was dropped for **Masa Yamamoto** will end up bringing a win. For better or worse, I think this year's team depends on those two. What about Byon? They'll get 1st place again next year, of course, that's after they win the **Central League** championship and the championship game.

Fig. 2 Blog document *d*[1]

Association words and fields for blog *d*. The items in bold and underlined in Fig. 2 are association words.

The calculated total for the associated words w for field *f* ="baseball" as shown in Table 2 is as follows.

POINT (*f,w*)*FREQ(*d,w*) = (40×1) + (30×2) + (30×2) + (30×1) + (40×1) + (35×2) = 300 (6 association words)

From the results shown above, we determine that "baseball" is the priority field for blog *d*.

Table 2: Results of extracting the field association words from document *d*.

W	*f*	POINT (*f,w*)	FREQ (*d,w*)	POINT (*f,w*) *FREQ (*d,w*)	Total
"Year-end grand prize"	"Prize /Lottery"	60	1	60 × 1	60
"Blog"	"Internet"	10	1	10 × 1	10
"Dragons"	"Baseball"	40	1	40 × 1	40
"Central League"	"Baseball"	30	2	30 × 2	60
"Japan Series"	"Baseball"	30	2	30 × 2	60
"Exhibition match"	"Baseball"	30	1	30 × 1	30
"Pitching staff"	"Baseball"	40	1	40 × 1	40
"Masa Yamamoto"	"Baseball"	35	2	35 × 2	70

[1] Goo Blog, http://blog.goo.ne.jp/.

2.3. Sensibility

Sensibility is considered to be among the abilities of human perception. It is the impression that is directly acquired via the sense organs. It is the subjective recognition that reflects an individual's characteristics and experiences. Sensibility is a sensory response that is difficult to express objectively in words, just like logical reasoning or decision-making logic.

However, words are one effective means of communicating Sensibility. Therefore, it is believed that it is possible to understand the intent of humans by understanding the sensory information that is included in words. In the Yoshinari method [14], sensory expressions, which can get imbued with a sense of an individual's personal opinion, are extracted as Sensibility from electronic documents. The Sensibility expressions that are found in free writing are grouped into the following three categories: Sensibility, emotion, and evaluation (Table 3). In this paper, we consolidate the Sensibility category and emotion category and refer to them as Sensibility. Each category is shown in Table 3 and marked with << >>.

Sensibility *S* shows the Sensibility presented in Table 3, analyzes each sentence in document *d*, and Sensibility *S* = SENSIBILITY(*x*,*y*) is defined from the modification relation (*x*,*y*). Then, when SENSIBILITY(*x*,*y*) is determined, *x* is the target word, and *y* is the evaluation expression. For example, the modification relation for "I won the lottery" is *x* = "lottery", *y* = "I won". Then, SENSIBILITY(*x*,*y*) is defined as <<happy>>. S_FREQ (*d*,*S*) is the frequency that Sensibility *S* appears in document *d*.

Algorithm 2. To determine Sensibility *S* and the priority Sensibility of blog *d*.

Step 1: Analyze each sentence in document *d*, and define Sensibility *S* = SENSIBILITY(x,y) from the modification relation (x,y).

Step 2: For Sensibility *S* = SENSIBILITY(*x*,*y*), calculate S_FREQ(*d*,*S*), and that total is determined to be the weight S_WEIGHT(*S*,*F*) of Sensibility *S* for document *d*.
Step 3: Identify priority SENSIBILITY(*x*,*y*) from multiple Sensibility *S*.

The Sensibility *S* that has the highest appearance frequency S_FREQ (*d*,*S*) is used as the priority Sensibility for blog *d*.

Table 4: Results of extracting the field association words from document *d*.

Target word *x*	Evaluation expression *y*	*Sensibility S*	Appearance frequency S_FREQ (*d*,*S*)	S_WEIGHT (*S*,*F*)
"Match"	"Win"	<<Fun>>	3	3
"Giants"	"Defeated"	<<Anticipation>>	1	1
"Two wins"	"Over"	<<Dissatisfaction>>	2	2
"Four wins"	"One loss"	<<Interesting>>	1	1

Table 4 shows the Sensibility results for blog document *d*. Among the multiple Sensibility, Sensibility *S* = <<fun>> and has a appearance frequency S_FREQ (*d*,*S*) of 3. This is the highest appearance frequency compared with that of other Sensibility, so we determine that <<fun>> is the priority Sensibility for blog.

Table 3: Results of extracting the field association words from document *d*.

Category	Types of *Sensibility* expressions					
Evaluation	Good		Bad			
Emotion	<<Happiness>>	<<Enjoyment>>	No emotion	<<Sorrow>>	<<Anger>>	<<Fear>>
Sensibility	<<Happy>>	<<Anticipation>>	<<Like>>	<<Sad>>	<<Bad fortune>>	<<Fear>>
	<<Good fortune>>	<<Fun>>	<<Praise>>	<<Bad news>>	<<Dissatisfaction>>	<<Worry>>
	<<Happiness>>	<<Funny>>		<<Resentment>>	<<Uneasiness>>	

2.4. Sensibility

Here, we present the proposed summary document generation method.

1. Using Algorithm 1, determine the field *f* that is the priority field and the find associated words *w*.
2. Using Algorithm 2, determine the Sensibility and find the priority Sensibility.
3. The summary document is generated using the three attributes that were acquired in steps 1 and 2: the field *f*, the field association word *w*, and the Sensibility *S*.

Table 5 presents a summary document that uses the field, field association word, and Sensibility, which were determined for blog d. Summary Document 1 can be generated by using the combination of the three attributes *f* = "baseball", *w* = 'match', and *S* = <<fun>>. Similarly, Summary Documents 2 and 3 can be generated using the same combinations. The method of generating summary documents shown in Table 5 is explained in further detail in Section 3.

Table 5: Summary document generation using attributes of fields, association words, and Sensibility

NO	Field *f*	Association word *w*	*Sensibility S*	Summary Document example
1	"Baseball"	'Match'	<<Fun>>	Fun is felt in relation to a baseball match.
2	"Education"	'Study'	<<Praise>>	Praise is felt in relation to educational studies.
3	"Love"	'Marriage'	<<Happiness>>	Happiness is felt in relation to a loving marriage.

3. Summary Document Generation from Attribute Grammar

3.1 Definition of Attribute Grammar

Attribute grammar [15] is a formal method to define the attributes related to formal grammar generation. In this paper, we introduce the attributes of fields, association words, and Sensibility, to attribute grammar. The attribute grammar is defined with the following four variables: G = (V_N,V_T,S,P). Here, V_N is a finite set of non-terminal symbols, V_T is a finite set of terminal symbols, S is the start symbol, and P is a finite set of production rules, which are given in the form of A(ATTRI)|COND|→α, where $A \in V_N, \alpha \in (V_N \cup V_T)^*$.

The start symbol $S \in V_N$ is the starting point for the generation of the selected sentence for that language. If a generation rule $A \rightarrow \alpha$ for the context-free grammar G exits, then symbol string : $\beta A \gamma$ ($\beta, \gamma \in (V_N \cup V_T)$, $A \in V_N$) will be overwritten with this generation rule.
Next, $\beta\alpha\gamma$ is derived directly from $\beta A\gamma$:

$$\beta A \gamma \underset{G}{\Rightarrow} \beta\alpha\gamma \quad (1)$$

Moreover, when

$$\alpha_1 \underset{G}{\Rightarrow} \alpha_2, \alpha_2 \underset{G}{\Rightarrow} \alpha_3, \ldots, \alpha_{n-1} \underset{G}{\Rightarrow} \alpha_n \left(\alpha_i \in (V_N \cup V_T)^*\right) \quad (2)$$

α_n Can be derived from α_1 :

$$\alpha_1 \overset{*}{\underset{G}{\Rightarrow}} \alpha_n \quad (3)$$

Hereinafter, when grammar G is clear, it will be abbreviated in the above notation. In addition, when symbol string $\alpha \in (V_N \cup V_T)^*$ is derived from the start symbol S, that is, when $S \overset{*}{\Rightarrow} \alpha$, α is referred to as a sentence form. Formal sentences that are created from only a terminal symbol are called sentences. Generally, sentence forms contain multiple no terminal symbols. The derivation that rewrites the leftmost (or rightmost) no terminal symbol is known as leftmost derivation (or rightmost derivation). The tree structure graph representation of the derivation process is known as a parse tree.

3.2 Definition of Attribute Grammar

In this research, sentences begin with *S*, *NP*, or other no terminal symbols. The subject words and object words that include attributes, such as the field's *f*, the association words *w*, and the Sensibility *S*, change depending on the noun phrase concept.

Therefore, we make the determination using the attributes at the *NP* generation stage. This procedure is formally defined in the following attribute grammar.

K-SET (*f*,*d*) is a keyword set that includes association words. S-SET (*d*) is a semantic co-occurrence (*x*,*y*) that

defines Sensibility and is a set that takes information as factors. SENSIBILITY(*x*,*y*) defines the Sensibility information for the semantic co-occurrence.

STRING (*f*) is a function that returns the field name as a text string for field *f*. The variable *P* is a particle in the attribute syntax definition. The variables *a*, *b*, and *c* are keyword elements other than association words and attributes. The variable *e* is a Sensibility expression. NC(*x*,*y*) is a group containing target word *x* and evaluation expression *y*.

Definition 2. Attribute grammar under generation conditions that use attributes.

1: < *S*(*f*,K-SET(*f*,*d*),S-SET(*d*))> → < *NP*(*f*,K-SET(*f*,*d*))>< *NP*(K-SET(*f*,*d*))>< PRED (S-SET(*d*))>
2: < NP(*f*,K-SET(*f*,*d*))> |*a*∈ K-SET(*f*,*d*) and *a*∈ F(*f*,*d*) and CONCEPT(*a*) have subject concepts ORGANIZATION and HUMAN |→<N P (*f*) ><N P(*a*)>
3: <N (*f*) >→STRING (*f*)/*"ORGANIZATION"*/
4: <P (*f*)>| (ORGANIZATION∈ CONCEPT (*f*) | >→"who is"
5: <N(*a*)>→*a* /* HUMAN */
6: <P (*a*)| (HUMAN∈ CONCEPT (β) | >→"the"
7: < NP (K-SET (*f*,*d*))> →< NP (K-SET(*f*,*d*))> < NP (K-SET(*f*,*d*))>
8: <NP (K-SET (*f*,*d*))> |b ∈ K-SET(*f*,*d*) and (TIME ∈ CONCEPT(*b*)) have object concepts TIME |→<N P(*b*)>
9: <N (*b*)>→*b*/* TIME */
10: <P(*b*)| (TIME∈ CONCEPT(*b*)|> →ε (empty symbol string)
11: < NP (K-SET (*f*,*d*))> |*c* ∈ K-SET(*f*,*d*) and PLACE∈ CONCEPT(*c*) |→<N P(*c*)>
12: <N(*c*)>→*c*/*"PLACE"*/
13: <P(*c*)|PLACE∈ CONCEPT(*c*) |>→"in"
14: < PRED (S-SET (*d*))>|(*x*,*y*) ∈ S-SET(*d*) and *e* = HAPPY = SENSIBILITY(*x*,*y*) |→<NC(*x*,*y*)><P(*x*,*y*)> <VP(*e*)>
15: <NC(*x*,*y*)> →*xy*
16: <P(*x*,*y*)> →"occurrence"
17: <V (*e*)>|*e*=HAPPY|→"Happy"
18: <P (*e*)>→"is"

Table 6 shows the concept names for the noun phrases of the subject words and object words that contain attributes

Table6: Concept names for attributes that are included in the definition

Subject concepts (*a*)	Object concepts (*b*)	Object concepts (*c*)	PRED concepts (*e*)
Organization	Time	Place	Happy
Human	Object	Agent body	Glad
Place	Material	Scene	Safety
Animal	Condition	Cause	Expectation
Plant	Modality	Starting point	Sad
Address	Target	Tool	Anger
Concreteness	Unit	Means	Worry

such as field's *f*, association words *w*, and *Sensibility S*, from the above definition.

If the summary document comes from the field names of individuals or organizations, and attributes generated with association words, then it will be generated using the above formal definition of the attribute grammar. The following is a specific example of this.

Example 1: *f* = "Baseball", K-SET (*f*,*d*) = {*a* = "Matsui", *b* = "the day before yesterday", *c* = "New York"}, S-SET(*d*) = {(x,y)} = {("a homerun","hit")}, SENSIBILITY(*x*,*y*) = *e* = <<fun>>, shows the generation for the previous summary.

Here, the fact that the K-SET (*f*,*d*) keyword "Matsui" is underlined indicates that it is an association word. For the field name *f* = "baseball", as defined above, STRING (*f*) is a function that returns the text string "baseball". The rule number used in the derivation is written first, as is done in the following example. C1, C2, C3, and C4 each show the generation conditions for the attributes. Figure 3 shows the parse tree for Example 1.

Table 7 shows the list of the production rule of the sentence generation rule proposes in the present study.

Table 7: List of generation rule of example 1

Syntactic rule	C1	→	|*a* = "Matsui"∈ K-SET (*f*,*d*) and(*a*)∈ F(*f*,*d*) and (HUMAN∈ CONCEPT(*a*)) |,
	C2	→	|*b* = "the day before yesterday"∈ K-SET (*f*,*d*) and (TIME∈ CONCEPT(*b*)) |,
	C3	→	|*c* = "New York"∈ K-SET (*f*,*d*) and PLACE∈ CONCEPT(*c*) |,
	C4	→	|(*x*,*y*)∈ S-SET(*d*) and *e* = <<fun>> = SENSIBILITY(*x*,*y*) |,
	N (*f*)	→	STRING (*f*) = "Baseball"
	N(a)	→	*a* = "Matsui"
	N (b)	→	*b* = "the day before yesterday"
	N(*c*)	→	*c* = "New York"
Syntactic rule	NC(x,y)	→	*x* = "a homerun" and *y* = "hit"
	P (*f*)	→	P1 = "who is"
	P(*a*)	→	P2＝"the"
	P(*b*)	→	P3 = ε
	P(*c*)	→	P4 = "in"
	P(x,y)	→	P5 = "occurrence"
	V (e)	→	"fun"
	P(*b*)		P6 = "isn't it"

The following shows the generation rules for example 1.

Rule 1: < S(*f*,K-SET(*f*,*d*),S-SET(*d*))> →<NP(*f*,K-SET(*f*,*d*))> <NP(K-SET(*f*,*d*)) < PRED (S-SET(*d*))>
Rule 2: < NP(*f*,K-SET(*f*,*d*)) |C1|>→<N P (*f*)><N P(*a*)>
Rule 3: <N (*f*) >→STRING (*f*)/*"Baseball"*/
Rule 4: <P (*f*)>| C1 | >→"who is"
Rule 5: <N(*a*)>→*a* /*Matsui*/
Rule 6: <P (*a*)| C1 | >→"the"
Rule 7: < NP (K-SET (*f*,*d*)) > →< NP (K-SET(*f*,*d*))> < NP (K-SET(*f*,*d*))>
Rule 8: < NP (K-SET (f,d)) |C2| >→<N P(b)>
Rule 9: <N (b)>→b/*the day before yesterday*/
Rule 10: <P(*b*)| C2|> →ε (empty symbol string)
Rule 11: < NP (K-SET (*f*,*d*)) | C3|> →<N P(*c*)>
Rule 12: <N(*c*)>→c/*"New York"*/
Rule 13: <P(c)| C3 |>→"in"
Rule 14: < PRED (S-SET (*d*)| C4| >→<NC(*x*,*y*) |><P(*x*,*y*)> < VP(δ)>
Rule 15: <NC(*x*,*y*)> →"a homerun" and "hit"
Rule 16: <P(*x*,*y*)> →"occurrence"
Rule 17: <V (e)>|e =HAPPY|→"fun"
Rule 18: <P (*e*)>→"it's"

Output generation summary sentence:
["It's a happy occurrence that Matsui, who is a baseball player; hit a homerun the day before yesterday in New York, isn't it?"]

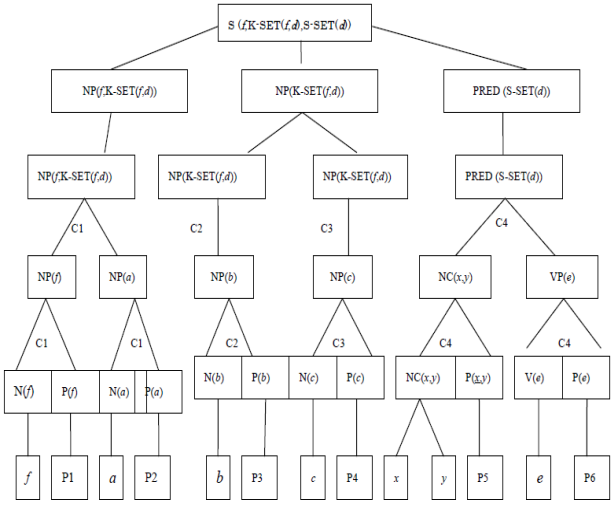

Fig. 3 Parse tree for example 1.

4. Experiment and Evaluation

4.1 Experiment Data

We used 400 randomly selected blogs across 9 categories from Live door Blog [16] , Goo Blog[17] , Blog People[18] and a blog search and ranking site[19] for the experiment. Table 8 below shows the fields, association words, sensibility, and keyword information that were acquired from each category which are necessary to

generate the summary document. It also includes detailed information on the size and number of documents in each category. In the experiment, categories such as "Anime", "Sports", "Politics", "Law", "Society", "Culture", "Economy", "Movie", and "Music" were selected and the following conditions were set as a criterion. The information is shown in Table 8.

1. It becomes the same name as the category as a result of a specific field of the blog, and it is written in the meaning that the content of the blog is the same as the category.
2. It is written in the meaning that the content of the blog is different from the category though it became a name different from the category as a result of a specific field of the blog.

The relevance ratio is used as a standard to evaluate the accuracy of the blog that meets the above-mentioned requirement. The calculating formula of the relevance ratio is as follows. The expression of the relevance ratio of each condition of the above-mentioned 1 and 2 will be shown, and it is shown relevance ratio 2 in the expressions of conditions of relevance ratios 1 and 2 in the expression of the condition of one.

$$\text{Relevance } 1 = \frac{C}{A} \times 100\% \quad \text{And} \quad \text{Relevance } 2 = \frac{C}{B} \times 100\% \tag{4}$$

A : Number of blogs that field and category of specified blog become the same names.
B : Number of blogs that field and category of specified blog become different names.
C : Number of blogs narrowed on condition of 1 and 2.

Table8: Outcome of an experiment of blog that meets requirement of selection

Category	A	B	C	Relevance 1 (%)	Relevance 2 (%)
Anime	64	65	45	70.3	69.2
Sports	75	71	60	80	84.5
Politics	60	65	45	75	69.2
Law	64	62	40	62.5	64.5
Society	80	82	60	75	73.1
Culture	70	73	53	75.7	72.6
Movies	62	65	47	75.8	72.3
Music	64	68	45	70.3	66.1
Economy	78	82	60	76.9	73.1
Total	617	633	455	73.7	71.8

4.2 Experiment Results and Evaluation

4.2.1. Summary Document Evaluation from Fields and Sensibility

Table 9 shows the experimental results. The experiments were performed under the standard that, for 400 sentences in a summary document, if all the attributes such as the fields, association words, and Sensibility were generated then the result was considered to be correct. If more than two were generated, then the result was deemed partially correct. However, if only one was generated, then it was incorrect. As a result of conducting the experiments under this standard, there were 355 correct results and the accuracy rate was 88.7%. There were 10 partially correct results, giving an accuracy rate of 91.2% if the partially correct items are included. The formula for calculating the accuracy rate is shown in Eq. (5).

Accuracy Rate (%) = Number of correct sentences / Total summary documents x 100%. (5)

The experiments were performed on blogs from a variety of categories. As a result, regardless of whether partially correct items were included or not, we were able to achieve an accuracy rate greater than 88%. Thus, we believe that this is an effective method for all categories of documents.

4.2.2. Evaluating the Summary Document for Readability and Meaning Integrity

We evaluated the readability and meaning integrity of 100 sentences that were randomly extracted from the 400 documents that were generated in this experiment. Our evaluation consisted of two parts. For the readability evaluation we checked whether the generated summary document used correct Japanese (Evaluation 1). For the meaning integrity evaluation, we compared the generated summary document with the original document and checked whether the meaning had been preserved (Evaluation 2). For the readability evaluation, we used four categories: A (extremely easy to read), B (easy to read), C (slightly easy to read), and D (difficult to read). For the meaning integrity experiment we used the following four categories: A (extremely easy to understand), B (easy to understand), C (can be understood slightly), and D (cannot be understood). These categories were used to evaluate the 100 sentences.

Table9: Results of experiment on summary document accuracy using the proposed method (unit: items/%).

Category	Summary sentences	Correct summaries	Partially correct summaries	Incorrect summaries	Summary accuracy rate[a] (%)	Summary accuracy rate[b] (%)
Anime	30	25	1	4	83.3	86.6
Sports	60	55	2	3	91.6	95
Politics	40	35	1	4	87.5	90
Law	35	30	1	4	85.7	88.5
Society	60	55	1	4	91.6	93.3
Culture	50	45	1	4	90	92
Movies	45	40	1	4	88.8	91.1
Music	30	24	1	5	80	83.3
Economy	50	46	1	3	92	94
Total	400	355	10	35	88.7	91.2

[a]Excluding partially correct summaries.
[b]Including partially correct summaries.

Table10: Evaluation of the readability and meaning integrity of the summary documents (units: number of sentences).

Evaluation category	A	B	C	D	Accuracy rate
Readability (Evaluation 1)	51	34	10	5	85%
Meaning Integrity (Evaluation 2)	56	30	10	4	86%

The results are shown in Table 10.

The evaluation results show that more than half of the 100 sentences from the summarized document were “easy to read”, and “extremely easy to read”. Moreover, the evaluation results for the readability and meaning integrity were 85% and 86%, respectively; thus, we were able to confirm the effectiveness of the overall summary document.

4.3 Comparison with Previous Research

In order to confirm the effectiveness of our method, we conducted a comparative evaluation with previous research. The previous research that we chose for our comparison was the summary research reported in Ref. [10] by Original method. This research uses introduces impressive expressions for newspapers and their measurements are applied to the NMF method. We used this research method to perform experiments on 100 blog sentences. According to the results considering the fields and Sensibility in the summary document, the summary accuracy rate was 84.5%. Table 11 shows the evaluation results for readability and meaning integrity.

Table11: Readability and meaning integrity results from Original method.

Evaluation category	A	B	C	D	Accuracy rate
Readability (Evaluation 1)	41	40	11	8	81%
Meaning Integrity (Evaluation 2)	44	38	10	8	82%

Table12: Comparison between results of Original method and proposed method.

Comparison item	Summary accuracy	Readability	Meaning integrity
Original methods	84.5%	81%	82%
proposed method	88.7%	85%	86%

Table 12 brings together the evaluation results from Table 9 for our research, and the previous research results from Table 11. From the results in Table 12, we can see that the summary accuracy rate of our method is 4% higher than the results of the previous method. We can also see that the readability and meaning integrity evaluations were each 4% higher.

4.4 Evaluation System and Simulation Results of Weight

In this paper, ROUGE scores were computed by running ROUGE-1.5.5 [20] with no stemming and no removal of stop words. The input file implemented so that scores of systems and humans could be compared. ROUGE evaluation systems are used to compute precision by using ROUGE_N which represents precision between generated summary of the proposed system and manual summary. Let n be the length of the *n-gram*, $gram_n$ is the maximum number of n-gram in the generated summary and $Count_n(gram_n)$ is a set of manual summary.

$$ROUGE\_N = \frac{\sum_{S\in\{manual\,summary\}} \sum_{gram_n\in S} Count_n(gram_n)}{\sum_{S\in\{manual\,summary\}} \sum_{gram_n\in S} Count\,(gram_n)} \quad (6)$$

In the system, five automatic evaluation methods are prepared in the ROUGE evaluation system ROUGE_N, ROUGE_L, ROUGE_W as follows:

1)ROUGE_N: *N-gram co-occurrence statistics* which is a precision between a generated summary of the proposed system and manual summary.
2)ROUGE_L: *Longest Common Subsequence (LCS)* which compares similarity between two documents in automatic summarization evaluation.
3) ROUGE_W: *Weighted Longest Common Subsequence* (*WLCS*) which is called a weight algorithm to assign different credit to consecutive in sequence matches.

In this paper, the summarization performances for the four weighting schemes are evaluated [21].

No weight Wgt(j,i) = tji (7)

Ordinary weight Wgt(j,i) = tji*log(N/n(i)) (8)

Binary weight Wgt(j,i) =1 if term i appears at least once in the sentence;

Otherwise Wgt(j,i) = 0 (9)

Modified binary weight Wgt(j,i) = tji*log(N/n(i))
If term i appears at least once in the sentence;
Otherwise Wgt(j,i) = 0. (10)

Fig. 4 show the ROUGE-L comparison results for the four weighting schemes on the proposed method. The ROUGE-W comparison results are shown in Fig. 5, and the ROUGE-SU comparison results are shown in Fig. 6, respectively. Moreover, the No weight for the recall evaluation results showed the best performance among the measures of ROUGE-L, ROUGE-W, and ROUGE-SU whereas the ordinary weight showed the best performance in the measure of ROUGE-SU. In the precision evaluation results, the binary weight showed the best performance among all ROUGE measures.

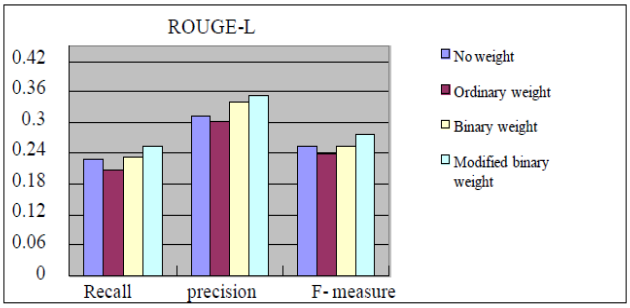

Fig. 4 Comparison using weighting schemes of ROUGE-L.

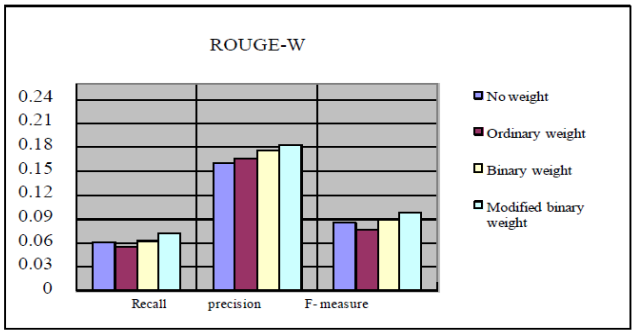

Fig. 5 Comparison using weighting schemes of ROUGE- W

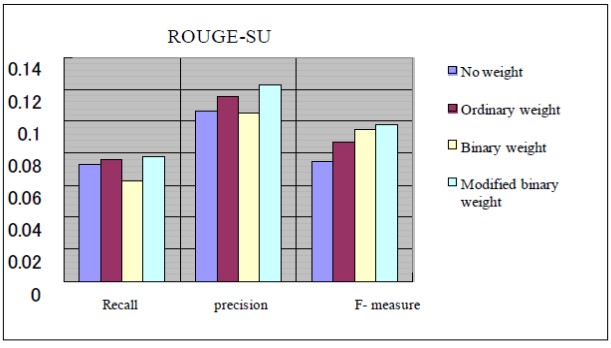

Fig. 6 Comparison using weighting schemes of ROUGE-SU

In the F-measure evaluation results, the modified binary showed the best performance among all ROUGE measures

4.5 Speed Experiments

The proposed system has been developed Linux Ubuntu 10.04.2 LTS 64bit and 8 CPU of Intel Xeon W3520 (2.67 GHz) with 6 GB main memory. Table 14 shows the speeding up of the computation of the NMF method for both original and proposed methods. The result in Table 13 is a measurement of the field, the sensibility analysis, and the time to generate the summary. The average of the document length of Table 9 becomes 20 lines because 35 characters become one lines in the actual experiment, and the analytical summary time of one document is 0.1 seconds. In 86 characters or less, 0.13 seconds have passed since it became 24 lines the summary time that hangs to it. Therefore, it has been understood that the proposal technique is very practicable.

In this experiment, one line consisted of 35 characters. Therefore, the average document length in Table 9 was 20 lines, and the analysis summary time for one document was 0.1 s. The maximum of 86 characters is 24 lines; thus, the summary time required was 0.12 s. Hence, we found that the proposed method is sufficiently practical.

4.6 Observations

When the number of sentences used in the calculations was few for some of the blog documents that we collected, we found that the fields, association words, and Sensibility were sometimes not able to be correctly acquired. This caused incorrect results to occur more frequently than usual.

In order to examine this, Fig.7 shows the relationship between the number of characters in the document and the accuracy rate of the summary document. From this figure, we can see that when there are 50 or more characters then the accuracy rate of the sentences exceeds 80%. From these results, we find that by using documents with more than 50 characters it is possible to improve the accuracy rate of the summary document.

In regards to the summary document evaluation, in the evaluations of readability and meaning integrity, there were summary documents that were difficult to read and also sentences that were incomprehensible. We believe that this issue can be improved if we increase the amount of field information and Sensibility information included in the summary document for these types of sentences.

Table13: Sentence analysis and summary document generation time. (Line)/ (ms)

Sentence number	10	15	20	25	30	35	40	45	50
Sentence processing time	50	80	100	130	150	200	250	300	350

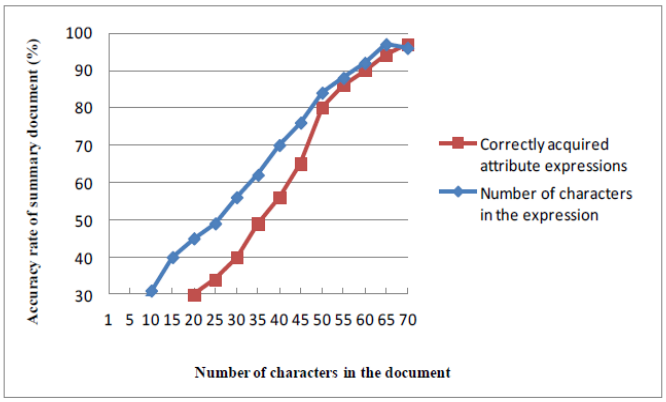

Fig. 7 Experimental results for accuracy rate of summary document

4.7 Evaluating Generation Using Attributes Grammar

In this research, we were able to use attribute grammar to generate summary documents. When we created the summary documents, we focused on the field, keywords, and Sensibility. Therefore, in order to do syntactic parsing correctly, the existence of attribute grammar was a necessity. Moreover, by keeping the field-focused noun phrases and Sensibility -focused verb phrases in the correct order, and by using attribute grammar, we were able to correctly generate the sentences.

5. Conclusions

In this paper, we have presented a method in which we use field association words and Sensibility, to create summary documents using attributes from the text information, such as fields, keywords, and Sensibility, which was taken from blogs. For the materials used to generate summary documents, first we used field

association words with the data acquired from the blog, and determined the blog's field. Then, we performed Sensibility analysis of the emotions of the people that appear in the contents of the blog and determined the Sensibility. For the summary document, once all three attributes were prepared, by using the attribute grammar, we established a formal and efficient generation technology.

From the results of our experiments using information from 400 blogs, we achieved a summary accuracy rate of 88.7% and found the proposed method to be effective.
In the future, we would like to research the translation of the generated summary document into other languages. We would also like to examine a method for setting the sentence construction rules such that they are created automatically or semi-automatically.

References

[1] T. M. Chang, W. F. Hsiao, "A hybrid approach to automatic text summarization", IEEE International Conference, 2008, pp. 65–70.

[2] L.Hennig, W.Umbrath, R.Wetzker, "An ontology-based approach to text summarization", IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 2008, Vol. 3, pp. 291–294.

[3] S. F. Liang, S. Devlin, J. Tait, "Investigating sentence weighting components for automatic summarization", Information Processing & Management, 2007, Vol.43, No.1, pp. 146–153.

[4] V. R. Uzeda, T. Pardo, M. Nunes, "Evaluation of automatic text summarization methods based on rhetorical structure theory", Eight International Conference on Intelligent Systems Design and Applications, 2008, Vol.2, pp. 389–394.

[5] A. Chongsuntornsri, O. Sornil, "An automatic Thai text summarization using topic sensitive page rank", International Symposium on Communications and Information Technologies, 2006, pp. 547–552.

[6] G. Erkan, D. R. Radev, L.Rank, "graph-based lexical centrality as salience in text summarization", J. Artif. Intell. Res, 2004, pp. 457–479.

[7] H. Zha, "Generic summarization and key phrase extraction using mutual reinforcement principle and sentence clustering", In Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval, 2002, pp. 113–120.

[8] J.Y.Yeh, H.R.Ke, "Text summarization using a trainable summarizer and latent semantic analysis", Information Processing & Management, 2005, Vol.41, No.1, pp. 75–95.

[9] L. H. Reeve, H. Han, "The use of domain-specific concepts in biomedical text summarization", Information Processing & Management, 2007, Vol.43, No.6, pp.1765–1776.

[10] A.Ubul, El.Atlam, H. Kitagawa, M. Fuketa, K. Morita, J. Aoe, "An Efficient Method of Summarizing Documents Using Impression Measurements", An Efficient Method of Summarizing Documents Using Impression Measurements, 2013, Vol.32, No.2, pp.371-391.

[11] H.J.Lee, S.Park, D.kim, "Automatic generic document summarization based on non-negative matrix factorization", Information Processing & Management, 2009, Vol.45, No.1, pp. 20-34.

[12] E.-S. Atlam, M. Fuketa, K. Morita and J. Aoe, "Document similarity measurement using field association term", Information Processing & Management, 2003, Vol.39, No.6, pp.809–824.

[13] E.-S. Atlam, G. Elmarhomy, M. Fuketa, K. Morita and J. Aoe, "Automatic building of new field association word candidates using search engine", Information Processing & Management, 2006, Vol.42, No.4, pp.951–962.

[14] T. Yoshinari, E.-S. Atlam, M. Fuketa, K. Morita and J. Aoe, "Automatic acquisition for sensibility knowledge using co-occurrence relation", International Journal of Computing and Technology, 2003, Vol.33,No.3, pp.218–225.

[15] F. Neven,J. V. den Bussche, "Expressiveness of structured document query languages based on attribute grammars", JACM, 2002, Vol.49,No.1, pp. 56–100.

[16] Livedoor Blog, http://blog.livedoor.com/.

[17] Goo Blog, http://blog.goo.ne.jp/.

[18] BlogPeople, http://www.blogpeople.net/.

[19] Blogger, http://blogger.bz/index.shtml.

[20] C.Y.Lin, "ROUGE: A package for automatic evaluation of summaries", In Proceedings of workshop on text summarization branches out, post-conference workshop of ACL, 2004.

[21] Y.Gong, X.Liu, "Generic text summarization using relevance measure and latent semantic analysis", In Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrival, 2001.pp.19–25.

Abdunabi Ubul received his B. Sc. degree in economics and Management information from Xinjiang University, China in 2004. He has received his M. Sc. degree from Department of Economics, Faculty of Integrated Arts and Sciences, University Of Tokushima, Japan in 2008. Received his Ph. D. degree from Department of Information Science and Intelligent Systems. University Of Tokushima, Japan in 2012. His research interests include information retrieval, natural language processing and document processing.

Hidekazu Kakei received his B.Eng. and M.Eng. Degrees in architecture from Nagoya University, Japan, in 1988 and 1990 respectively, and his Ph.D. in architecture in Kobe University in 2007. Since 2003 he has been an Assoc. Prof. in the Institute of Socio-Arts and Sciences, Tokushima University, Japan. His research interests include applying ICT to spatial and environmental design. He is a member of Architectural Institute of Japan and the Institute of Electronics, Information and Communication Engineers.

Jun-ichi Aoe received his B. Sc. and M. Sc. degrees in electronic engineering from the University of Tokushima, Japan, in 1974 and 1976, respectively, and his Ph. D. degree in communication engineering from the University of Osaka, Japan in 1980. Since 1976 he has been with the University of Tokushima. He is currently a Professor in the Department of Information Science&Intelligent Systems, Tokushima University, Japan. His research interests include design of an automatic selection method of key search algorithms based on expert knowledge bases, natural language processing, a shift-search strategy for interleaved LR parsing, robust method for understanding NL interface commands in an intelligent command interpreter, and trie compaction algorithms for large key sets. He is the editor of the Computer Algorithm Series of the IEEE Computer Society