Inductive Learning of Rules for Information Extraction

Takuto Tsukahara, Kenji Araki, *Member, IEEE,* and Koji Tochinai

Abstract-- **There are many information extraction systems that help to save time for reading a lot of documents. The information extraction is the method to extract important information from a document. Generally conventional information extraction methods need to prepare many rules for extracting important information. The pattern of extracted information has to be fixed. Therefore they are effective for the limited fields when it is obvious what kind of information a user wants. However, it is not effective when a user reads the documents of various fields. In this paper, we propose an information extraction method for Japanese documents using Inductive Learning. The system learns what kind of information a user needs and the system gets several rules for information extraction from the correct answers given by a user. The system uses two kinds of rules to learn the user's wants. One is the rule to decide the important sentences. And the other is the rule to extract the important words. Using these rules, the system can adapt to a user dynamically. When user's interest changes to other topics, the system can extract information a user wants. The system is able to realize to extract important information from the documents of the various fields. In this paper, we explain how to extract important information and describe the detail of two rules for information extraction. And we evaluate the effectiveness of our proposed method. The recall and the precision of the rules to decide the important sentences is over 80% after the learning progresses. Therefore the rule to decide the important sentences is effective for the various fields. However there are some problems in the rules to extract the important words. The problems are the variety of the output patterns and the method to apply the rules. We consider the causes and describe the solution.**

I. INTRODUCTION

Recently, the opportunity to read documents on a computer is increasing with development of the Internet. The number of documents that one can get to read is over the limit that a human has. There are many information extraction systems[1][2] that help to save time for reading many documents. The information extraction is the method to extract important information from documents. Generally conventional information extraction methods need to prepare many rules for extracting important information. They extract information using the pattern matching. The pattern matching is simple method compared with the method of summary. Conventional information extraction methods are applied when it is obvious what kind of information a user wants. These are effective to the documents on limited fields. However, it is not effective when a user reads documents on the various fields because when the field is changed, a user has to prepare new rules for new field. This work is difficult to the user without expertise on information extraction and the user has to take much time for the work.

In this paper, we propose the information extraction method for Japanese documents on various fields using Inductive Learning[3]. Inductive Learning is to get the rules that inhere in the example. We define the process that a common part and a different part are extracted recursively as Inductive Learning. Using Inductive Learning, our proposed method predicts and extracts the important information that a user needs from a document. The system learns what kind of information a user needs and gets several rules for information extraction from the correct answers given by a user. A user has to give the correct answers to the system for learning. It is easier than preparing new rules for extraction because a user has only to choose the words he wants from documents without expertise.

We aim at realization of the information extraction to various documents with our proposed method. In this paper, we describe the effectiveness of this proposal method with performance evaluation experiment.

II. OUTLINE OF OUR PROPOSED METHOD

The system based on our proposed method uses two kinds of rules. One is to decide the important sentences in a document. The sentence containing the words a user wants is defined as the important sentence. And the other is to extract important words from the important sentences and output the words. We explain these two kinds of rules in the next chapter.

The overview of our method is shown in Figure 1. At first, a morphological analysis is carried out to input document using morphological analysis tool ChaSen[4] for Japanese. Next, the important sentences are chosen using the rules. And important information is extracted from the important sentences using the other rules. Those two processes are also explained in the next chapter. Through these two processes, the extracted information is outputted. If an error were occurred in the processes, the user would have to proofread the results. At the time, learning of the two kinds of rules is carried out. In this process, two kinds of the rules are registered in the rule dictionary. At the end, feedback of the two rules is carried out. A degree of priority of the rules used by mistake is lowered. The rule that the correct answer rate is low is deleted from the rule dictionary.

This work is partially supported by the Grants from the Government subsidy for aiding scientific researches (No.14658097) of the Ministry of Education, Culture, Sports, Science and Technology of Japan.

Takuto Tsukahara and Kenji Araki are with the Graduate School of Engineering, Hokkaido University, Kita 13 Nishi 8, Kita-ku, Sapporo, Hokkaido, 060-8628, Japan.
(e-mail: {tukahara, araki}@media.eng.hokudai.ac.jp).

Koji Tochinai is with the Graduate School of Business Administration, Hokkai Gakuen University, Asahimachi 4-1-40, Toyohira-ku, Sapporo, Hokkaido, 062-8605, Japan. (e-mail: tochinai@econ.hokkai-s-u.ac.jp)

ISBN: 1-86467-114-9

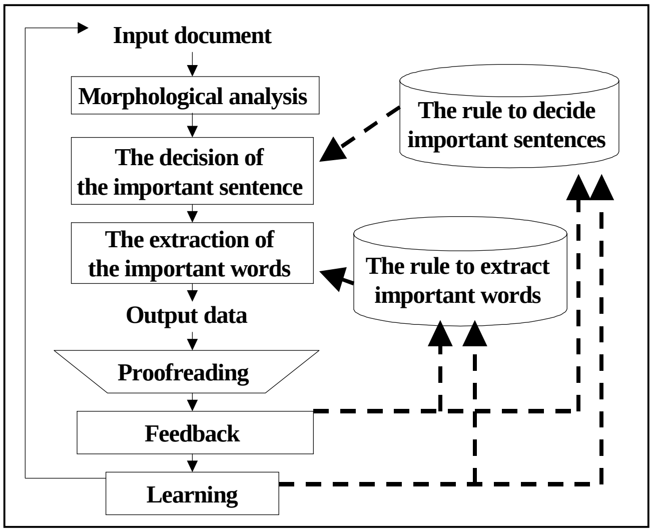
Input document
Morphological analysis
The decision of the important sentence
The extraction of the important words
Output data
Proofreading
Feedback
Learning
The rule to decide important sentences
The rule to extract important words

Figure 1 Process

III. TWO KINDS OF RULES

A. *The rule to decide the important sentences*

This system uses the rule to decide the important sentences. This type of rule consists of 9 elements. Those elements show the various contents of a sentence. Each element has the number that expresses the state of the element. The detail of 9 elements is shown in Figure 2. We define the number string in which those 9 numbers are expressed as the parameters of the sentence. The system is able to decide the important sentences using these parameters and the rule dictionary.

The parameters of the important sentences are registered in the rule dictionary. The examples of the rule are shown in Figure 3. In the rule dictionary, the rule that took the common part between two rules is contained. For example, in Figure 3, “1x1x0xxxx” shows the common part between “111200120” and “101001001”. “x” is the different part. And each parameter in the rule dictionary has correct answer rate. It expresses the precision of the parameter. The correct answer rate is used to apply the rules.

Next, we explain how to decide the important sentences. Each parameter of the sentences in an input document is compared with the rules in the dictionary. If all number of the elements in a rule agrees with the parameter of the input sentences, the rule is used for the calculation to decide a degree of importance. For example, if the parameter of the input sentence is “101000101” as shown in Figure 3, the rules “1x1x0xxxx” and “xx1xx01xx” in the rule dictionary are used. The calculation to decide a degree of importance is carried out in each sentence of the input documents. The calculation to decide a degree of importance is shown as equation (1). When a sentence agrees with the rule of which the correct answer rate is high, the degree of importance of the sentence is high through this calculation. And the rule that contains no “x” has more influence than the rule that contains many “x”. The sentences with high degree of importance are determined as important sentences.

Element 1: The position of the sentence
1: The first sentence 2: A previous part
3: The middle 4: A latter part
Element 2: Paragraph
1: The first paragraph 2: The final paragraph
3: Others 4: No paragraph
Element 3: The position of the paragraph
1: The first sentence 2: A previous part
3: The middle 4: A latter part
Element 4: Connection
1: Normal 2: Reverse 3: Addition 4: Reworded
5: Illustration 6:Reason 7: Comparison
8: Conversion 9: Other
Element 5: Items
1: Here 2:No
Element 6: Date expression
1: Here 2:No
Element 7: The type of the sentence
1: Guess 2: Request 0: Other
Element 8: The score of the keyword
0: No 1: 0.33 2: 0.67 3: 1 4: 1.5
5: 2 6: 2.5 7: 3 8: 3.5 9: 4.0
Element 9: The importance of Noun
0: No 1: 1 2: 2 3: 3 4: 4
5: 5 6: 6 7: 7 8: 8 9: 9

Figure 2 Element of the sentence

The rule dictionary to decide important sentences

The rule	The correct answer rate
111200120	0.8
101001001	0.5
1x1x0xxxx	0.3333
231110111	1.0
xx1xx01xx	0.75

The parameter of the input sentence

101000101

Used rules

1x1x0xxxx and xx1xx01xx

Figure 3 The rule to decide the important sentences

$$A = \frac{\sum Rate * (9 - X) * Dignity}{Number} \qquad (1)$$

A: The degree of importance is shown
Rate: The correct answer rate of the rule
X: The number of “x” in the rule
Dignity: The sum of dignity
Number: The number of the used rule

The important sentence
Puro yakyu se ri-gu deha 16 nichi, kyojin ga toukyou do-mu de hanshin to taisenshi, 1 – 4 de kanpai shita.
(The game of the professional baseball was held on 16, and “Kyojin” faced “Hanshin” in Tokyo dome, and a score was 1-4.)

The correct answer given by a user
Nichiji : 16 nichi (Date : 16th)
Taisen ka-do: kyojin – hanshin (Card : team’s name “kyojin” – team’s name “hanshin”)
Kekka : 1 – 4 (Result : 1 – 4)

The rule to extract the important words
Input : *1*[Noun-number] *6*[Noun-number] *nichi*[Noun-connection] *kyojin*[Noun-proper noun- organization] *hanshin*[Noun-proper noun- organization]*1*[Noun-number] –[Sign-general] *4*[Noun-number]
Output : *nichiji:1*[Noun-number] *6*[Noun-number] *nichi*[Noun-connection] */ taisenka-do : kyojin*[Noun-proper noun-organization] –[Sign-general] *hanshin*[Noun-proper noun- organization] */ kekka : 1*[Noun-number] –[Sign-general] *4*[Noun-number]

Figure 4 The rule to extract the important words

The important sentence of the input document
1[Noun-number] *8*[Noun-number] *nichi*[Noun-connection] *no*[Particle] *puro*[Noun-general] *yakyu*[Noun-general] *ha* [Particle] *hito*[Noun-number] *siai*[Noun-sahen] *okonawa*[Verb] *re*[Verb] ,[Sign-punctuation] *yakuruto*[Noun-proper noun-organization] *ga*[Particle] *hanshin*[Noun-proper noun-organization] *wo*[Particle] *2*[Noun-number] –[Sign-general] *0*[Noun-number] *de*[Particle] *kudashi*[Verb] *ta*[Auxiliary] .[Sign-punctuation marks]
(Professional baseball on 18 is 1 game, and “Yakuruto” beat “Hanshin” 2 – 0.)

The rule to extract the important words (input)
Input : *1*[Noun-number] *6*[Noun-number] *nichi*[Noun-connection] *kyojin*[Noun-proper noun- organization] *hanshin*[Noun-proper noun- organization]*1*[Noun-number] –[Sign-general] *4*[Noun-number]

The calculation shown as equation (2)
B = (3 + 2 + 3 + 2 + 3 + 2 + 3 + 2) / 8 = 2.5

The rule to extract the important words (output)
Output : *nichiji:1*[Noun-number] *6*[Noun-number] *nichi*[Noun-connection] */ taisenka-do : kyojin*[Noun-proper noun-organization] –[Sign-general] *hanshin*[Noun-proper noun- organization] */ kekka : 1*[Noun-number] –[Sign-general] *4*[Noun-number]

The result of information extraction
Nichiji : 18 nichi (Date : 16th)
Taisen ka-do: yakuruto – hanshin (Card : team’s name “Yakuruto” – team’s name “Hanshin”)
Kekka : 2 – 0 (Result : 2 – 0)

Figure 5 The extraction of the important words

A. *The rule to extract the important word*

After the important sentences are decided, the required words in the important sentences are extracted according to the form that a user desires. We use the other rules to extract the important words. These rules consist of two elements. One is the sequence of the words for input and the other is for output. This is registered in the rule dictionary from the correct answer. The sequence of the words for output is equal to the correct answer given by a user. It is the sequence of the words a user wants following the order that a user requests. And the sequence of the words for input is the sequence arranged the order that they exist in the sentence.

An example of the rule is shown in Figure 4. This important sentence is sports news about baseball game. If a user wants “date”, “card” and “result”, a user needs to give the correct answer shown in Figure 4 to the system. The sequence of these words is the rule for output. And the rule for input is the sequence of the words that exist in the important sentence.

Figure 5 shows how to extract the important words. The important sentences are compared with all the rules in the dictionary. If the same sequences of the words or the words of the same part of speech in the rule for input exist in the input sentence, the rule is used. In Figure 5, the underlined words in the important sentence agree with the words in the rule for input. If the rules exist more than two, the calculation to decide a degree of importance is carried out. The calculation is shown as equation (2). If a word in the important sentence is the same with a word in the rule for input, we define the score is 3. If the part of speech is the same, we define the score is 1. If the classification of the part of speech is the same, we define the score is 2. For example, about “*1*[Noun-number]” in the important sentence in Figure5, the score is 3. About “*yakuruto* [Noun-proper noun-organization]”, the score is 2. And the average of scores is calculated. The rule keeping the highest score is used. After the rule is decided, the sequence of the word is outputted as the result of information extraction according to the rule for output. When there is no rule to extract important words, the important sentence is outputted. This is a solution for the early stages of learning.

$$B = \frac{\sum Score}{Number} \qquad (2)$$

Score: The score of the rule as follows:
3: All corresponds.
2: The classification of the part of speech corresponds.
1: A part of speech corresponds.
Number: The number of a part of speech

IV. Performance evaluation experiment

The experiment for evaluating the performance of our proposed method was carried out. 150 documents about sports news[5] were used in the experiment. Using this system based on our proposed method, we carried out information extraction about 150 documents. We evaluate this system. The number of total sentences is 1,138. These sports news were chosen at random. There are various documents in them, for example baseball, soccer, ski, sumo and so on. The detail of the experiment is described as mentioned below.

A user made the correct answers about 150 documents. The correct answers are information a user wants. At first, the rule dictionary was empty because we want to know the process of a user model for learning. Information extraction was carried out about 150 documents. After information extraction about each document is finished, a user proofread and the system got the rules.

We evaluate two kinds of rules. The standards for evaluation of the rule to decide important sentences are the recall and the precision. The recall and the precision are shown as equation (3) and (4). The recall expresses how many important sentences are extracted. The precision expresses how many sentences in extracted sentences are correct. This result is shown in Figure 6.

$$\mathrm{Re}\,call(\%) = \frac{Number1}{Number2} * 100 \qquad (3)$$

$$\mathrm{Pr}\,ecision(\%) = \frac{Number1}{Number3} * 100 \qquad (4)$$

Number1: The number of the important sentences extracted properly.
Number2: The number of the important sentences of the correct answer.
Number3: The number of the sentences extracted.

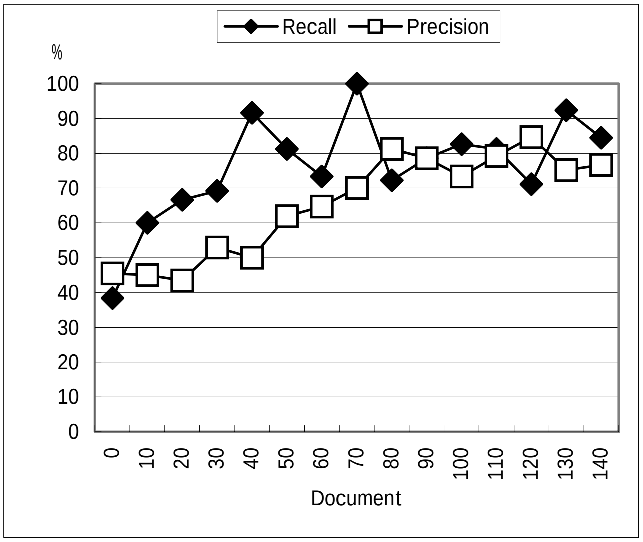

Figure 6 The recall and the precision of the rule to decide the important sentences

And the classification of the extracted result using the rule to extract the important words is shown in Figure 7. In Figure 7, we classify the documents under three topics “A”, “B” and “C”. “A” expresses that the correct answer given by a user is written in the pattern of output and there is the same pattern of output in the rule dictionary. “B” expresses that the correct answer given by a user is written in the pattern of output and there is not the same pattern of output in the rule dictionary. “C” expresses that the correct answer given by a user is sentence as it is. We classify the result about these three types documents. The document that the correct answer is extracted is “Correct”. The document that the wrong words are extracted is “Wrong”. The document that there is no proper rule in the dictionary and the important sentence is outputted is “No”. The correct answer rate is shown as equation (5).

$$Rate(\%) = \frac{(13+33)}{(69+42)} * 100 = 41.4 \qquad (5)$$

	Correct	**Wrong**	**No**	**Total**
A	**13**	21	35	**69**
B	-	6	33	39
C	-	9	**33**	**42**
Total	13	36	101	150

A: The correct answer is in the pattern.
The same pattern exists in the dictionary
B: The correct answer is in the pattern.
The same pattern does not exist in the dictionary.
C: The correct answer is sentence as it is.
The same pattern does not exist in the dictionary.

Figure 7 The classification of the extracted result

V. CONSIDERATION

A. *The rule to decide the important sentences*

In Figure 6, the recall and the precision are low in early stage for learning. However they increase as learning progresses. The recall is more than 80% and the precision is about 80% in the documents of the portion after the learning of 100 documents. This result shows the effectiveness of this rule to decide the important sentences. This rule is effective about the documents of various fields. However it is necessary to increase the recall. Because the shortage of important information in an answer from the system is more problematic than the answer included several extra information. We have to improve the decision of the important sentences to increase the recall.

B. *The rule to extract the important words*

We cannot get enough result about the rule to extract the important words. In Figure 7, this system could extract correct answers about 13 documents. About 39 documents, the rule dictionary has no rule to extract the correct answer because the pattern of output appears at first. Although these 39 documents are removed, it is not good result. The correct answer is extracted about 13 documents and the correct sentences given by a user are extracted about 33 documents. Even if we consider that these two cases are the correct answers, the correct answer rate is 41.4%.

One of the main causes of this result is the variety of the output patterns. All the experiment data were sports news. However the contents were the various topics of many sports as Figure 8. For example, the result of the game such as soccer, the topic of the player's transfer and the result of the individual event such as skiing and judo. For its reason the amount of learning data for a pattern of output was too small. We should prepare for some similar data to learn a pattern of output. We got some correct answer about it because there were many documents about the result of soccer. Therefore similar data are necessary to learn a pattern of output.

Soccer J-league (12), **Soccer in Europe** (22),
Soccer World Cup (7), **Soccer etc** (11),
Baseball in Japan (18), **Baseball Major League** (6),
Baseball etc (9), **Player’s Transfer** (15),
Skiing (9), **Basketball** (6), **Track and field** (5),
Rugby (4), **American football** (4), **Sumo** (3),
Judo (2), **Skates** (2), **etc** (15)

(): The number of the documents

Figure 8 The kinds of the documents

Though the important sentences were decided precisely, we could not extract the important words according to the pattern same as the correct answer given by a user. This cause is that the wrong rule was applied when there was no proper rule. When it is certain that the rule to extract the important words is proper, the rule should be applied. However, when the rule is unreliable, the system should output the important sentence as it is. This compromise plan is effective in early stage for learning. We must establish the condition to choose the rule for extraction. The rule with high value of the equation (2) is applicable.

Next, we consider the cause that the system used the wrong rule. There are some causes as follows:

(1) The classification of the part of speech

We used the calculation as equation (2) in this experiment. In this method, if a part of speech agrees between the word in the input rule and in the important sentence, the rule is usable.

Though the purpose of this method is to give the rule a generality, it was the cause that the wrong rules were used. When there was no proper rule in early stage for learning, the wrong rule with low value of the equation (2) was applied. It is necessary to give the rule a generality, and we have to establish the condition to apply the rule. For example, when the value of the equation (2) is more than 2.0, the rule is applied. The opportunities that the rules are applied may decrease. However it is desirable that the important sentence is outputted as it is when there is no proper rule because of the high precision of the rule to decide the important sentences.

(2) The document contains the plural similar information

The system extracted only one answer when there was the plural similar information in the document. For example, when there are the results of two games in the document about baseball, the system cannot extract both of them using the rule for the result of one game. The rule for the result of one game is different from the rule for the results of two games. This problem can be solved by the following method. The method has two kinds of the rules. The examples of them are shown in Figure 9. One expresses that a user wants to extract “date”, “card” and “result” from the document about baseball. The other expresses that what kind of word is proper in “date”, “card” and “result”. By this method, the system can get more data for learning when the common part like “date” in the various documents. And we expect that learning is faster.

The rule for the outline

Yakyuu = Nichiji / Taisen ka-do / Kekka
(Baseball = Date / Card / Result)

The rule for the details

Nichiji = 1[Noun-number] *6*[Noun-number] *nichi* [Noun-connection]

Taisen ka-do = kyojin[Noun-proper noun- organization] –[Sign-general] *hanshin*[Noun-proper noun - organization]

Kekka = 1[Noun-number] –[Sign-general] *4*[Noun -number]

Figure 9 New rules for solution

(3) Difference in the turn of the words

Though there was the same pattern of output in the rule dictionary, the correct answer was not extracted because of difference in the turn of the words. This problem can be solved as the data increases. However we have to solve this problem to increase the precision in early stage for learning. We will try to solve this problem by the generalization of the rules.

We will try to solve these problems and increase the precision of the rule to extract the important words in future. We have to devise how to learn from a small quantity of data. And we will try to increase the correct answer rate in early stage for learning by the method to output important sentence as it is when the rule is unreliable.

VI. Conclusion

We proposed the information extraction method using Inductive Learning. This system based on our proposed method learns what kind of information a user needs and adapts to a user dynamically. It is possible that information extraction to the documents of various fields by this method. The experiment was carried out and we considered the effectiveness of the two rules. The recall and the precision of the rules to decide the important sentences increase as learning progresses. The recall is more than 80% and the precision is about 80% in the documents of the portion after the learning of 100 documents. We could describe that this rule is effective about the documents of various fields. On the other hand, there are some problems in the rules to extract the important words. The problems are the variety of the output patterns and how to apply the rules. We will improve our system to solve these problems by the method stated in the consideration. And we’d like to show the effectiveness of the rule to extract the important words with evaluation experiment in future.

References

[1] Pazlenza, M. T. (ed.): Information Extraction, Springer-Verlag. Lecture Notes in Artificial Intelligence, Rome (1997).

[2] Grishman, R. and Sundhelm, B: Message Understanding Conference-6: A Brief History, The 16th International Conference on Computational Linguistics (COLING-96).

[3] K. Araki and K. Tochinai, "Effectiveness of Natural Language Processing Method Using Inductive Learning", Proceedings of the IASTED International Conference ARTIFICIAL INTELLIGENCE AND SOFT COMPUTING, pp.295-300, May, 2001, Cancun, Mexico.

[4] D. Matsumoto, "Morphological analysis system ChaSen version 2.0 manual" NAIST Technical Report, NAIST-IS-TR99008, April 1999.

[5] http://www.asahi.com.