Assistant Tool for Concealing Personal Information in Text

● Tomoya Iwakura ● Seishi Okamoto ● Kunio Matsui

(Manuscript Received November 20, 2006)

This paper describes an assistant tool for concealing personal information in text. This is an important procedure for protecting privacy when public documents are disclosed, preventing accidental leaks of personal information, and other purposes. However, finding personal information in text is very time-consuming and labor-intensive. To make it easier to conceal personal information, we have developed a graphical user interface (GUI) tool that extracts candidate personal information in text, indicates candidate personal information using different colors according to its class, and creates rules for extracting personal information from text, including annotations of personal information. In one experiment, our GUI tool enabled users to conceal personal names in Japanese text about three times faster than when the task was done without candidate personal information.

1. Introduction

In Japan, the protection of personal information is now a major problem for the government and enterprises when they disclose their administrative documents, and there have recently been frequent leaks of this type of information. The Act on Access to Information Held by Administrative Organs[1] became fully effective in 1999 to provide the right to request the disclosure of administrative documents. However, administrative documents contain personal information, and Japanese law states that personal information in these documents must not be publicly disclosed. Furthermore, a Japanese law that protects personal data in the documents of enterprises became fully effective in April 2005,[2] and as a result, Japanese companies now have to pay more attention to disclosure risks.

Documents containing personal information can be safely released by first concealing the information. However, this takes a long time and is labor-intensive: the most crucial task being to find the personal information. Workers have to find personal information by reading the entire text very carefully because the positions of personal information are not fixed in unprocessed text written in natural language.

In this paper, we present an assistant GUI tool for concealing personal information in text. This tool has the following features.

1) Extraction of personal information: The tool extracts several types of proper nouns and numeric expressions together with their classes, for example, personal names and the names of locations, organizations, and dates, as personal information candidates in text.
2) Indication of candidate personal information: The tool indicates candidate personal information using colors according to its class with an extraction summary, and users can conceal the candidates by just clicking them.
3) Creation of rules for extracting personal information: The tool creates rules for extracting personal information from text,

including annotations of personal information to be concealed.

Section 2 of this paper describes a process for concealing personal information in text. Section 3 describes our assistant GUI tool for concealing personal information. Section 4 reports experimental results on concealing personal information in a pseudo Japanese court decision. Section 5 shows how our approach differs from other approaches, and Section 6 presents our conclusion.

2. Concealing personal information in text

Text data includes several types of person identifying information, for example, people's names, dates of birth, and addresses. To protect privacy, we must find this information and transform it into non-identifying information.

For example, in Japan, courts do not disclose court decisions as is because they include personal data and other confidential information.[2] Therefore, before a court decision can be disclosed, all the personal information it contains must be removed.

The following example shows how text data can be anonymized or concealed.

1) Original text: "The accused man, Mr. Yamada, was released after agreeing to testify against the others. Mr. Yamada testified that…"
2) Anonymized text: "The accused man, Mr. X, was released after agreeing to testify against the others. Mr. X said that…"

In this example, even though the personal name has been anonymized, we can still understand the text. Therefore, organizations can conceal personal information in their documents and then safely disclose it. However, it takes a long time to conceal personal information because most of the work has to be done manually. To make this work easier, we propose the use of tools that automatically find and indicate personal information that needs to be concealed. Although there are currently no tools that can completely extract personal information, we think that these assistance tools can greatly reduce the amount of manual work that must be done.

3. Our assistant tool for concealing personal information

This section describes our assistant tool for concealing personal information in text using a natural language processing technology called Named Entity (NE) extraction.

3.1 Extracting personal information by using context information

One of the problems in concealing personal information in text is that we have to correctly determine the meanings of words from their contexts.

For example, most location names in Japan, such as the names of prefectures, cities, and so on, are also Japanese family names. Therefore, to determine whether these names are family names, we must consider the contexts in which they appear.

To correctly find personal information in text, we apply NE extraction technologies. NE extraction aims to identify word chunks such as proper nouns and numerical expressions and then classify them as, for example, persons, locations, organizations, or dates. **Table 1** shows some NE examples. Our personal information extraction methods are outlined below.

NE representation in text: Because NEs consist of one or more words, we employ a method that classifies words into NE labels by using context information. We used Start/End (SE) representation, which uses five tags — S, B, I, E, and O — to represent word chunks.[3] "S" means that the current word is a chunk consisting of only one token. "B" indicates the start of a chunk consisting of more than one word. "E" indicates the end of a chunk consisting of more than one word. "I" indicates the inside of a chunk consisting of more than two words. "O" indicates the

Table 1
NE examples.

Numeric expression				
NE class	DATE	MONEY	PERCENT	TIME
Example	May 5th	200 JPY	100%	10 PM

Proper noun				
NE class	ARTIFACT	LOCATION	ORGANIZATION	PERSON
Example	Nobel Prize in Chemistry	Japan	Fujitsu	Jorge White

text outside of a chunk.

We use these five tags for NE labels to represent NEs in text.

For example, the person names and the text outside of the person names in:

"..., Mr. Michael W. White said to Mr. Brown..."

are represented as:

"..., Mr./O Michael/B-PERSON W./I-PERSON White/E-PERSON said/O to/O Mr./O Brown/S-PERSON..."

"B-PERSON" indicates the start of a person name, "I-PERSON" indicates the inside of a person name, "E-PERSON" indicates the end of a person name, and "S-PERSON" indicates a person name consisting of only one word.

Classifying words into NE classes by using context information: To classify words into NE labels, we use information about the two adjacent words as context in addition to the current word information. We use the following word information.

1) Words: Written Asian languages such as Japanese and Chinese have no word boundaries. Therefore, to segment words in Japanese sentences, we use a morphological analyzer and the NE extractors classify the segmented words into their proper classes.
2) Part of speech (POS) tags: We use the POS tags of words tagged by the morphological analyzer used for segmenting words.
3) Character types: We use various types for Japanese characters, for example, "hiragana," "katakana," and "Chinese letter;" "uppercase alphabet;" "digit;" and "sign" and combinations of these types.
4) Dictionaries: We use dictionaries that indicate personal information such as job titles and lists of addresses, if available. In this experiment, we employed a dictionary created from text by using several NE extractors.[4)]

For example, to classify "W." in the above sentence, the word information "Mr.", "Michael", "White" and "said" are used as context information to discriminate the NE class label of "W.".

Classifying characters into NE classes by using context information: Each Japanese NE consists of one or more words or includes a substring of a word. This is because Japanese words are not separated by spaces as in English and word segmentation depends on what types of dictionaries are used. For example, the "訪米" ("visit USA") in

田中使節団は訪米 (Tanaka mission party visit USA.)

does not match with LOCATION "米" ("USA") because this sentence is tokenized as:

"田中 (Tanaka) 使節 (mission) 団 (party) は (particle) 訪米 (visited USA)",

where indicates a word boundary.

To solve this problem, we use an NE extraction algorithm based on character-unit chunking[5)] after classifying words into NE classes. To classify characters into NE labels, we use information about the previous and following characters as context in addition to the current character information.

1) Characters and words: Words include the characters within the current word and its two adjacent characters. Characters are expressed with position identifiers to indicate where the characters appear in words. We use B, I, E, and S, which is the same as the SE representation.
2) POS tags: POS tags are annotated into words by a morphological analyzer. As with characters, POS tags are expressed with position identifiers in SE representation.
3) Character types: We use kanji, hiragana, katakana, numbers, lowercase and uppercase letters, and other characters.
4) NE labels of words: We use NE labels of words by stacking.[6] These labels are also expressed with position identifiers in SE representation.
5) NE labels of preceding extraction results: We use NE labels of the previous character and classify characters into NE labels from the end to beginning of a sentence.

Each character is classified into NE labels represented by IOB2[5] representation, which uses three tags: B, I, and O. "B" indicates the start of a chunk. "I" indicates the inside of a chunk consisting of more than two words. "O" indicates the outside of a chunk. The above example therefore is represented as:

"田/ORGANIZATION-B 中/ORGANIZATION-I 使/ORGANIZATION-I 節 団/ORGANIZATION-I は/O 訪/O 米/LOCATION-B",

where "ORGANIZATION-B" and "ORGANIZTION-I" indicate the beginning and inside of an organization name, respectively. "LOCATION-B" indicates the beginning of a location, and "O" indicates the outside of the organization and location names.

3.2 Indicating candidate personal information

One of the hardest procedures in concealing personal information in text is finding it. Because the positions of personal information in unprocessed text is not fixed, workers have to read an entire text to find all the personal information it contains. Therefore, to make it easier to find personal information, our GUI tool finds and indicates candidate personal information as follows.

1) Indicating candidates: Our tool indicates candidate personal information using colors according to its class, so users can easily find personal information by looking for the appropriate colors. The left window in **Figure 1** shows a snapshot of our tool. The red characters in bold font, 福岡史明 as an example, indicate personal names, and the black background of these characters indicates the candidate personal information that the mouse cursor is pointing to. Users of our tool can conceal a candidate by just clicking it. "[PERSON] (2)" to "[PERSON] (5)" in the figure indicate concealed person names.
2) Creating an extraction summary: Our tool creates a summary of the extraction consisting of the candidate personal information together with its surrounding words. This summary enables workers to conceal personal information without reading an entire text. The right window in Figure 1 shows a snapshot of a summary created by our tool. The user can conceal personal information in the text by clicking the candidates listed in the summary.

Our tool also provides users with shortcut keys for modifying incorrect extractions. **Figure 2** shows a snapshot of a modification for an incorrect extraction. The characters with the black background, 川崎次郎 in the Figure, indicate the incorrect extraction that is selected by the mouse. For example, if a user defines "P + Alt" as an annotating person tag, the user can annotate person tags by selecting the places to be annotated with the mouse and pressing P + Alt. Our tool also provides shortcuts for deleting incorrect extractions.

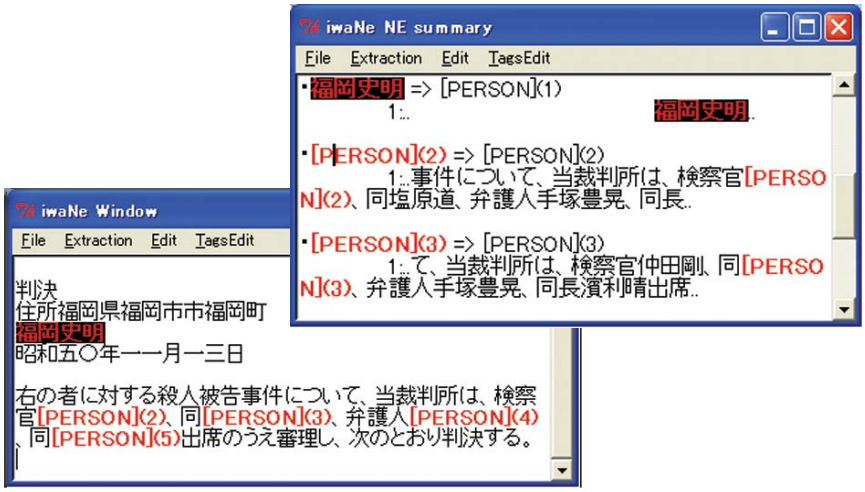

Figure 1
Snapshot of our tool: Left window shows entire text, and right window shows a summary. Concealed person names and a location name are represented by their classes and numbers. Black background indicates item selected by mouse pointer.

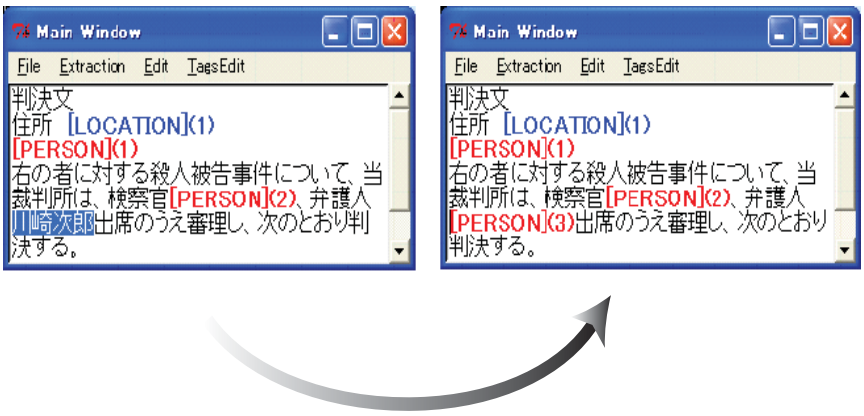

Figure 2
Snapshot of modification for incorrect extraction. Users can annotate tags of personal information by selecting the places of the information and pressing shortcut keys corresponding to personal information classes.

3.3 Creating rules for extracting personal information

There are two types of approaches for extracting personal information: handcrafted rule based approaches and machine learning based approaches.

If we use a handcrafted rule based approach for extracting personal information, we can revise the extraction by adding new rules or by modifying current rules. However, rule development is a skill that needs to be learned and rules take a long time to create. With an NE extraction approach based on machine learning, we can obtain extraction rules from examples such as text whose personal information has been concealed, but it is difficult to control the extraction behavior.

The two approaches have different benefits and drawbacks. However, we mainly employ a machine learning based approach because machine learning based approaches can create rules from new text that has been created for

disclosure in routine work and has had its personal information concealed.

The machine learning algorithm we use is a boosting algorithm. A detailed explanation of this algorithm is given in a paper elsewhere.[7] In our tool, the rule generator creates rules for extracting personal information from training data in which personal information is marked with class name tags bracketed with "<>". **Figure 3** shows an example of rule generation from training data created in a concealing process.

4. Experimental results

We evaluated the effectiveness of our tool by measuring the time needed to conceal personal names in a pseudo court decision for students studying law. The decision is contained in a 12 KB file and includes 45 personal names. The extraction rules we used were created by the tool from text that included 7818 personal names.

We evaluated our GUI tool's performance in terms of the following parameters:

Recall = NUM / number of person names to be extracted
Precision = NUM / number of person names extracted by our GUI tool
F-measure = 2 × Recall × Precision / (Recall + Precision),

where NUM is the number of person names correctly extracted. **Table 2** shows the extraction accuracy of our tool. As can be seen, our tool extracts person names at an F-measure of more than 90%.

Table 3 shows the results for concealing personal names with two subjects. The results show that our tool enabled the two subjects to conceal the names about 3.3 to 3.9 times faster compared to when they did the tasks manually.

5. Related work

There are three main approaches to extracting personal information. The first approach is dictionary based personal information extraction.[8] This approach has the advantage that users can enhance extractions. However, it cannot determine the meanings of words from their contexts.

The second approach is handcrafted rule based extraction.[9] This approach has the advantage that users can control the extraction behavior by creating and modifying rules. However, the creation of rules and the management of rules is very labor intensive.

The third approach — machine learning based extraction — is the one adopted by our tool.

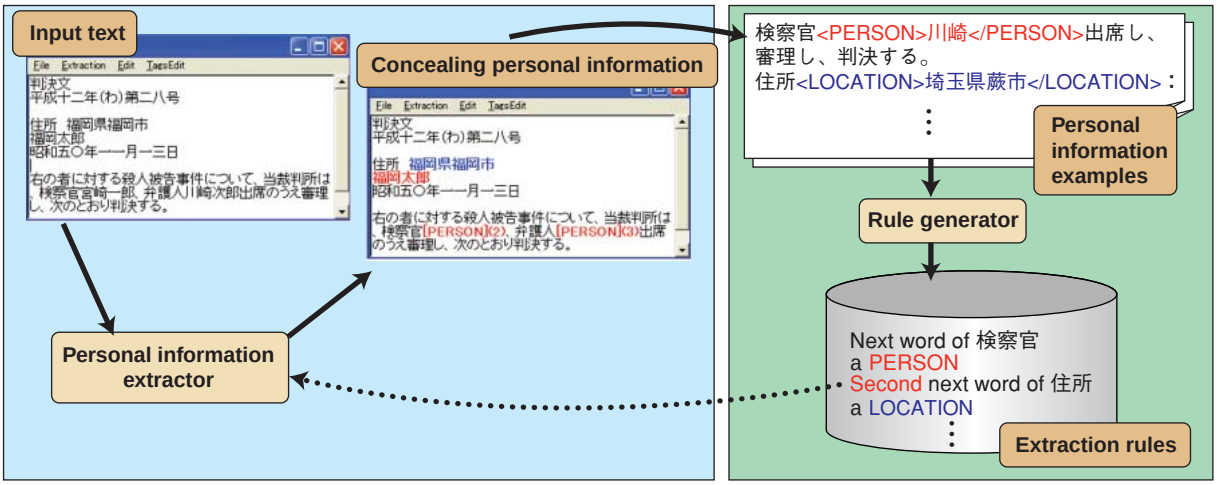

Figure 3
Snapshot of rule generation from data created in a concealing process. Left side shows a concealing process. Right side shows an example of rule generation.

Table 2
Extraction results of our GUI tool.

Number of person names	Precision (%)	Recall (%)	F-measure (%)
45	97.44 (37/38)	84.44 (37/45)	90.48

Table 3
Processing times for concealing personal information.

Subject	Manual operation (a)	With our tool (b)	Improvement ((a)/(b))
A	938 s	243 s	3.9
B	706 s	218 s	3.3

This approach uses machine-learning algorithms to create extraction rules from contextual information. Although it is difficult to control the extraction behavior in this approach, it has the advantage that users can create new rules by creating data annotated with personal information. Furthermore, this approach enables users to enhance the extraction performance by incorporating data that includes concealed personal information created in routine work.

Products for finding personal information are on sale in Japan.[10),11)] These products aim to assist in the management of files containing personal information on PCs so that accidental leaks of personal information can be prevented. Compared to these products, our GUI tool has two advantages. Firstly, it indicates candidate personal information to be concealed. Secondly, it provides a function for creating rules for extracting personal information from text whose personal information has been concealed.

6. Conclusion

This paper described our GUI tool for concealing personal information in text. This tool detects personal information by using a Japanese NE extractor for identifying personal information by using context information. It indicates candidate personal information, which the user can conceal by just clicking. We experimentally demonstrated that this tool reduces the time needed to conceal personal information in text by a factor of about three compared to when the work is done without candidate personal information.

Further improvements in concealing personal information will be realized by enhancing the performance of personal information extractors. For further improvement, future work should consider methods of combining approaches based on dictionaries, handcrafted rules, and machine learning algorithms.

References

1) Act on Access to Information Held by Administrative Organs (Act No. 42 of 1999). *http://www.soumu.go.jp/english/gyoukan/060516_03.html*
2) Web page for disclosed court decision. *http://www.courts.go.jp/saisinhanrei.html*
3) K. Uchimoto, Q. Ma, M. Murata, H. Ozaku, M. Utiyama, and H. Isahara: Named entity extraction based on a maximum entropy model and transformation rules. Proc. of the ACL 2000. 2000, p.326-335 .
4) T. Iwakura and S. Okamoto: Improving Named Entity Extraction Accuracy Using Unlabeled Data and Several Extractors. CICLing 2007 (Poster session) (to be published).
5) M. Asahara and Y. Matsumoto: Japanese named entity extraction with redundant morphological analysis. Proc. of HLT-NAACL 2003, 2003.
6) D. H. Wolpert: Stacked generalization. *Neural Networks*, **5**, 2, p.241-259 (1992).
7) R. E. Schapire and Y. Singer: BoosTexter: A boosting-based system for text. *Machine Learning*, **39** (2/3), p.135-168 (2000).
8) Dehenken Web page. (in Japanese). *http://www.dehenken.co.jp/products/products-03/products-kansalib01.html*
9) Y. Takemoto, T. Fukushima, and H. Yamada: A Japanese Named Entity Extraction System Based on Building a Large-scale and High-quality Dictionary and Pattern-matching Rules. (in Japanese), *IPSJ Journal*, **42**, 6, p.158-159 (2001).
10) Kensyutsu Meijin Web page. (in Japanese). *http://jp.fujitsu.com/group/tfl/services/kensyutsu/*
11) SecretBarrier Web page. (in Japanese). *http://www.infobarrier.com/secretbarrier/products/index.html*

Tomoya Iwakura, *Fujitsu Laboratories Ltd.*
Mr. Iwakura received the B.S. and M.S. degrees from Kyushu Institute of Technology in 2001 and 2003, respectively. He joined Fujitsu Laboratories Ltd. in 2003, where he has been engaged in research and development of natural language processing systems.

Seishi Okamoto, *Fujitsu Laboratories Ltd.*
Mr. Okamoto received the B.S., M.S., and PhD degrees from Kyushu University in 1989, 1991, and 1998, respectively. He joined Fujitsu Laboratories Ltd. in 1991, where he has been engaged in research and development of intelligent systems and information retrieval systems. He is a member of the Information Processing Society of Japan and the Japanese Society for Artificial Intelligence. He is a visiting associate professor at the Japan Advanced Institute of Science and Technology. He received the Best Paper Award from the Union of Japanese Scientists and Engineers in 1997 and the Systemics, Cybernetics and Informatics Best Paper Award from the World Multi-conference in 2002.

Kunio Matsui, *Fujitsu Laboratories Ltd.*
Mr. Matsui received the B.S. degree from Shizuoka University in 1980 and PhD degree from Tokyo Institute of Technology in 2003. He joined Fujitsu Laboratories Ltd. in 1980, where he has been engaged in research and development of natural language processing systems, information retrieval systems, and knowledge management systems. He is an executive board member of the Information Processing Society of Japan. He is also a member of the ISO/TC37 Committee of the Japanese Standards Association.