社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2009－DPS－138 (7)
2009－CSEC－ 44 (7)
2009／3／5

プライバシー情報検知のための知識の準備と学習
－自然言語情報の開示制御技術 DCNL の実現（２）－

水谷桂子* 片岡春乃† 吉浦裕*

*電気通信大学大学院 電気通信学研究科

*〒182-8585 東京都調布市調布ヶ丘 1-5-1

†日本電信電話株式会社 NTT 情報流通プラットフォーム研究所

†〒180-8585 東京都武蔵野市緑町 3-9-11

概要 BlogやSNSからのプライバシーや機密に関わる情報漏洩が多発し，社会問題になっている．そこで，その対策として筆者らはプライバシー情報を検知し適切な表現に変換する自然言語情報の開示制御技術DCNLを開発している．自然言語の日記からプライバシー情報を検知するためには，システムが膨大な知識を準備する必要がある．また，複数の文章にまたがった検知が必要である．本論文では，格フレーム検索ツールを用いて5億文のテキストコーパスから知識を自動抽出する方式を提案する．また，各々の文章から検出されたプライバシー情報にスコアをつけ，複数文章にわたってスコアを加算する情報検知方法を提案する．

キーワード 情報漏洩，プライバシー，アクセス制御，人工知能，認知科学

Learning Knowledge for Detecting Private Information from Natural Language Sentences

Mizutani Keiko* Kataoka Haruno† Yoshiura Hiroshi*

* Graduate school of Electro Communications, The University of Electro-Communications

* 1-5-1 Chofugaoka, Chofu, Tokyo, 182-8585 Japan

† NTT Information Sharing Platform Laboratories, NTT Corporation

†3-9-11, Midori-cho Musashino-Shi, Tokyo 180-8585 Japan

Abstract We are developing Disclosure Control of Natural Language information (DCNL) to prevent disclosure of private information from Web media. To learn knowledge about privacy that is necessary for DCNL, this paper describes a method of using a case-query tool to extract privacy knowledge from corpus of a half billion sentences. It also describes a method for accumulative detection of privacy by accumulating sensitivity scores from individual sentences.

Keyword Privacy protection, Web security, Disclosure control, Access control, Natural language analysis

1. はじめに

近年，BlogやSNS(Social Networking Services)などのWeb上のコミュニケーションメディアが急速に普及してきた．これらのメディアは個人を対象とし，Webや携帯電話から簡単に利用できる．そのため多くの人々が参加し，ユーザ同士でのコミュニケーションが活性化してきた．特にSNSでは，コミュニケーションをする相手は友人関係にある場合が多い．これは，友人同士のユーザ間で相互にユーザページをリンクする仕組みがあるためである．そのためSNSでは，コミュニケーションがより親密になる．

一方，このようにコミュニケーションが親密になり，かつ気軽に投稿できるため，個人の発信する情報量が増えた．そのためプライバシー情報などの要注意情報の漏洩や誹謗中傷などの不適切な表現が問題となっている．

SNSサイトの中には閲覧者に対する日記やプロフィールの閲覧の制限をユーザが設定できるものもある．たとえばSNSサイトのmixi[1]においては，ユーザ同士の関わり方によって，日記ごとに公開範囲を設定できる．すべてのユーザが閲覧可能な「全体公開」，友人関係であるユーザ

[1] mixi は株式会社ミクシィの社名であり，かつ同社の運営しているソーシャル・ネットワーキング サービスの名称である．

のみが閲覧可能な「友人までの公開」などユーザ同士の関係によりアクセス制御をする．さらに，ユーザが選択した友人のみが閲覧できる「グループ公開」もある．このような選択的なアクセス制御により，プライベートな内容ならば「グループ公開」にするといった，ユーザ自身による要注意情報の保護が可能になった．

しかし，日記の内容は多種多様であるため，日記ごとに内容を丁寧にチェックし，閲覧者を細かくグループに分けて，閲覧者のアクセスを制御することはユーザにとって負担が大きい．また，要注意情報の見逃しなど，想定外の要注意情報の漏洩がありうる．さらに，日記単位でアクセスが制御されるため，公開されなかったユーザはその日記は何も閲覧できない．そのため，アクセスが制御されたユーザは，その日記に含まれる要注意情報以外のすべての情報も，閲覧できなくなる．これは，SNS上のコミュニケーションを過度に制限することになり，本来の面白さを失ってしまう．

我々は，Web上のコミュニケーションのための自然言語情報の開示制御DCNL(Disclosure Control of Natural Language information)[1][2][3]を提案し研究している．これは，ユーザがWebメディアに投稿した文章を自動的に解析し，要注意情報に該当する語句を言い換えるなどの開示制御を行うことで情報漏洩を未然に防ぐ．これにより，ユーザの負担になることなくメディア上でのコミュニケーションを安全に行えるようにし，かつ過度な制限を防ぐことができる．

本稿ではDCNLの実現に向け，要注意情報に関わる知識の準備と，複数文における自然言語テキストからの要注意情報検知について検討する．

2. DCNL(Disclosure Control of Natural Language information)

2.1. DCNL の要件

(1) Web メディアに投稿された文章を解析し，要注意情報を漏らす恐れのある語句について削除，または言い換える．

(2) 言葉の組み合わせにより生じる意味や間接的な意味を推測する．また，Web 検索による要注意情報の発覚も考慮する．

(3) 言い換えにあたって，文章の意味と面白さを維持する．

(4) ユーザへの負担を最小限に抑える．例えば，ユーザの要注意情報は出来るだけ自動的に収集するようにし，ユーザによる入力などの負担を少なくする．

(5) 要注意情報に関する知識の補完・学習を自動的に行う．

2.2. DCNL のシステム構成

図１にシステム構成図を示す．Web メディアにアクセスすると，閲覧者が認証され，ユーザとの関係が特定される．DCNL は関係クラスにより原文を言い換え安全な文章にする．始めに，「自然言語処理」が，原文の語句を認識する．次に，「要注意情報の検知」が「要注意情報に関する知識」を用いて，語句が要注意情報かを判断する．最後に「言い換え処理」が，文章の意味や面白さを維持しながら問題語句をユーザとの関係性に応じて言い換えるか削除し，最終的な文章を生成する．

2.3. プライバシー情報検知のための知識

以下，要注意情報のうちプライバシー情報の検知に限定して説明する．

DCNL では Web メディアに投稿された文章から，プライバシー情報を検知するために以下の３種類の知識を準備することが提案されている．

① NG ワード

ユーザそれぞれの具体的なプライバシー情報となる語句．たとえば，具体的な住所である「調布」や「小島町」などの地名である．これは，個人により異なるため，SNS のユーザプロファイルなどから一部を取得し，さらにユーザによる入力やチューニングが必要である．

② 種類語

検知するプライバシー情報が何か，プライバシー情報の種類を示す語句．たとえば，「住む」や「暮らす」という動詞で、これ

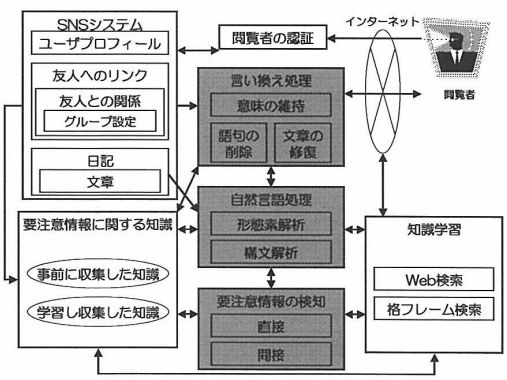

図１ DCNL システム構成

らが住所に関する【種類語】だと分かることで，その文章が住所に関するプライバシー情報を表していると推測できる．

③ 特徴語

接尾辞や接頭辞などの表現の特徴から意味を推測する語句．たとえば，「市」や「区」などで，それが住所に関する【特徴語】だと判断できればその語句の前後の語句が住所に関するものだと推測できる．

【NG ワード】がユーザ固有であるのに対し，基本的な【種類語】と【特徴語】はユーザ共通である．DCNL ではこれらの知識を組み合わせることによりプライバシー情報を検知する．

3. DCNL の実現に向けた課題

3.1. 課題の概略

DCNL の実現に向けプライバシー情報検知技術において検討すべき課題を示す．

まず，プライバシー情報の検知のための知識の準備である．【NG ワード】や【種類語】，【特徴語】は膨大かつ多種多様であり，すべてを人手で準備することは非常に難しい．そこで，これらの知識を自動的に作成し，さらに DCNL の使用に伴い知識を追加学習する手法を検討する．

次に，Web 上の知識を利用したプライバシー情報の抽出手法の検討がある．DCNL では自然言語を分析するために形態素解析を用いて原文の語句を認識するが，誤って認識される語句や認識できない未知語がある．このとき，Web 上の知識を用いて，語句の誤りの訂正や未知語の意味の推測を行う．

最後に，複数文章からの解析方法の検討を行う．日記は複数の文章からなる場合が多く，さらに日々投稿されることにより数多くの文章が蓄積される．そのため過去に検知したプライバシー情報を学習し，それを新たな検知に利用する．また 1 つの文章だけではプライバシー情報かが曖昧であっても，複数の文章を見ることでプライバシー情報だと判断できる場合もある．

本稿では，これらの課題のうち，プライバシー情報に関する知識の自動収集の手法と，複数文章からのプライバシー情報検知技術について述べる．

3.2. 事前知識の準備の課題

DCNL でプライバシー情報を検知するための知識は，ユーザ固有か共通かに大別できる．

住んでいる場所や所属は個人で異なるため，【NG ワード】はユーザ固有のものである．そのため，SNS のユーザプロファイルからの自動取得に加え，ユーザ自身によるチューニングが必要となる．そこで，ユーザの負担を減らすため，自動化によりユーザを支援する必要がある．

一方，基本的な【種類語】と【特徴語】は，ユーザに関わらず共通だといえる．たとえば，「住む」という語句は，誰にとっても住所に関する【種類語】になる．これらの知識は，共通知識としてシステム開発時に事前に準備するが，膨大かつ多種多様のため，それらすべてを人手のみで準備することは非常に難しい．そこで，システム開発者による知識の準備を支援するための自動化処理が必要となる．

3.3. 複数文章における解析の課題

SNS に投稿される文章は複数になる場合が多い．そのため，DCNL ではプライバシー情報検知のため複数文章にまたがる解析の必要がある．

そこで，男性ユーザの 1 年半の日記と女性ユーザの日記を分析し，複数文章からプライバシー情報が検知されるパターンを 2 つ見出した．

1 つめは，日記から検知し学習した新しいプライバシー情報の知識を，それ以降の日記の解析に利用することで，プライバシー情報の検知が可能になる場合である．たとえば，日記に『仙川に住んでいる』という文章があったとする．この文章では，「仙川」が地名かどうか分からない場合でも，【種類語】の「住む」から住所に関するプライバシー情報の可能性があるとして「仙川」を学習できる．さらに，後日の日記に『仙川のスーパーで買った』や『仙川には深夜に着いた』という文章が記述された場合，この 2 つの文章には【種類語】がないため「仙川」を住所に関するプライバシー情報とは判断できない．しかし，最初の文章から「仙川」を住所に関するプライバシー情報の可能性があると学習していれば，後の 2 つの日記でも住所に関するプライバシー情報だと判断できる．また，同じ語句を繰り返し抽出したことで，「仙川」がプライバシー情報である可能性は高いと判断できる．このとき，最初の日記がない場合は「仙川」をプライバシー情報だと学習していないため，「仙川」という語句が何度も投稿されるだけでは，繰り返し抽出される語句は他にも多くあるので，「仙川」をプライバシー情報とは検知できない．

2 つめは，複数の文章の組み合わせによりプ

ライバシー情報の曖昧性の解消が可能になる場合である．たとえば，ある日記に『小島町に住んでいる』という文章があるとき，「住む」という【種類語】から「小島町」が住所に関するプライバシー情報だと推測できる．しかし，「小島町」という地名の地域はいくつかあるため，プライバシー情報として曖昧である．しかし，他の文章に県名や市名など，地域を特定できる記述があれば曖昧性を解消することが可能になる．

以上より，複数文章の解析のために日記内の他の文章や過去の日記から抽出した知識を利用する必要がある．そこで，DCNL では，ユーザごとに抽出した知識を学習し，以後の検知に利用するための蓄積方法と，蓄積した知識を用いた検知方法の確立が必要である．

4. 事前知識の準備

ここでは，ユーザ共通の知識である【種類語】と【特徴語】の取得手法について述べる．

4.1. 知識の準備の方針

基本的な【種類語】と【特徴語】は，ユーザに関わらず共通である．そのため，Web 上の文章から抽出できる【種類語】や【特徴語】は，すべてのユーザに対応する．そこで，Web 上の文章をコーパスとして解析し【種類語】や【特徴語】を抽出し学習する．

人手により準備した【種類語】や【特徴語】と同じ格フレーム構造をもつ語句を Web 上のコーパスから抽出し，新たな【種類語】や【特徴語】として学習する．たとえば「住む」や「暮らす」などと同じ格フレーム構造を持つ他の語句を抽出できれば，その語句は住所に関する【種類語】である可能性がある．

そこで，京都大学が Web 上に公開している格フレーム検索ツール[4]を用いて，同じ格フレーム構造を持つ語句を抽出する方法を検討する．

4.2. 格フレーム検索ツール

格フレームに該当する語句を多数の文例から抽出するツールとして格フレーム検索が Web 上に公開されている．格フレーム検索は Web 上の 5 億の文をコーパスとしている．これは，ある動詞の格にどのような語句が該当するのか，また，ある語句を格として取りうる動詞には何があるのかを文例から抽出するものである．たとえば，「住む」という動詞に対して格フレーム検索を行うと『調布に住む』や『一人で住む』

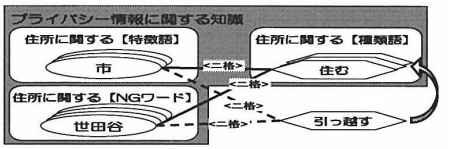

図2 格フレームの実現イメージ

のような文例から「住む」の<ニ格>として「調布」が，<デ格>として「一人」が抽出される．格フレーム検索は Web 上の文章がコーパスのため，SNS 上の文章に近く，DCNL における検知に利用するのに適していると考えられる．

4.3. 実現方式

(1) 格フレームを利用した【種類語】の準備

人手により準備した【特徴語】や【NG ワード】と，対応する【種類語】に対して格フレーム検索を行いそれぞれの格フレームの格を割出し，同じ格フレームの動詞を抽出する．この抽出された動詞は同じ種類のプライバシー情報の【種類語】となりうる．以下にその過程を示す．また，この実現方式のイメージを図 2 に示す．

Step1 知識として人手で準備した【特徴語】や【NG ワード】と，対応する【種類語】に対して格フレーム検索を行い，それぞれの格フレームの格を割出す．

Step2 【特徴語】や【種類語】に対する格フレーム検索の結果から，Step1 で割出した格の動詞を抽出する．

Step3 抽出された動詞が，新たな【種類語】となりうる語句か検討する．

たとえば，事前知識として，住所に関する【特徴語】の「市」や【NG ワード】の「世田谷」と，【種類語】の「住む」を保有している場合，「市」と「住む」，「世田谷」と「住む」に対する格フレーム検索を行うことで，それぞれの格フレームの格が<ニ格>だと分かる．さらに「市」や「世田谷」に対する格フレーム検索を行う．その結果を表 1 に示す．頻出上位 10 件ずつを抜粋した．検索結果より<ニ格>の動詞を抽出すると，両方の検索結果に含まれるものが「ある」

表1 格フレーム検索結果(左：市 右：世田谷)

格フレームID	格	頻度	格フレームID	格	頻度
ある:動1	二格	22993	住む:動1	二格	95
住む:動1	二格	7255	ある:動1	二格	66
生まれる:動1	二格	3395	いる:動1	二格	27
市:判1	無格	3241	住む:動1	ノ格	16
位置:動1	ノ格	2948	引っ越す:動1	二格	15
開催:動:P1	デ格	2553	生まれる:動1	二格	12
開く:動:P1	デ格	2338	実現:動944	二格	11
ある:動1	ノ格	2315	考える:動2	ノ格	8
開く:動:P1	ノ格	2071	語る:動13	ヲ格	8
なる:動1	二格	1947	来る:動1	二格	8

「生まれる」，片方のみに含まれるものが「引っ越す」「なる」「来る」「実現」「いる」．このうち「引っ越す」「生まれる」の<ニ格>は住所に関する【種類語】になる可能性がある．

以上より，抽出結果のすべてが種類語であるわけではないが格フレーム検索を利用することで新しい【種類語】の多様な候補を抽出できる．

(2) 格フレームを利用した【特徴語】の準備

すでに保有している【種類語】に対して格フレーム検索を行うことで，その【種類語】を格とする名詞の抽出ができる．この抽出された名詞は【特徴語】となりうる語句の可能性がある．以下にその過程を示す．

Step1 人手により準備した【種類語】と対応する【特徴語】に対して格フレーム検索を行い，その格を割出す．

Step2 【種類語】に対する格フレーム検索の結果から，Step1 で割出した格に該当する語句を抽出する．

Step3 抽出された語句が，新たな【特徴語】となりうる語句か検討する．

たとえば所属に関するプライバシー情報の知識として【種類語】の「通う」，【特徴語】の「大学」を保有しているとする．この「大学」と「通う」に対する格フレーム検索により，この格フレームの格が<ニ格>であると分かる．さらに，【種類語】である「通う」に対しての格フレーム検索を行うと，この動詞のそれぞれの格に対応する名詞を抽出できる．この結果から，<ニ格>に対応する名詞の頻出上位を見ると「大学」以外では「学校」「教室」「スクール」「高校」「ジム」などがある．これらは，所属に関する【特徴語】になりうる．

以上より，すべてが【特徴語】になるとは言い切れないが，格フレーム検索を利用することで新たな【特徴語】の候補を抽出できる．

5. 複数の文章からの情報の積み重ね

5.1. 方針

(1) 検知した結果に対するスコア

DCNL では【NG ワード】のみの場合，【NG ワード】と【種類語】の場合，【特徴語】と【種類語】の場合の３通りの組み合わせによりプライバシー情報を検知する．この組み合わせにより抽出される情報はプライバシー情報である確信度が異なる．たとえば【種類語】と【特徴語】の組み合わせから推測された場合より【NG ワード】が直接書かれている場合の方がプライバシー情報漏洩の危険性が高い．また，「調布に住む」の「調布」のように【種類語】と直接係り受けにある場合もプライバシー情報を表している可能性が高い．そこで検知した言葉の種類や係り受けに応じて検知したプライバシー情報にスコアを付加することで抽出された情報に重み付けを行う．これにより，スコアの高さに応じてユーザに危険性を示すことが可能になり，また，すべての情報がない場合でもある程度の検知が可能になる．

(2) 複数文章におけるスコア加算

複数文章からの情報を積み重ねていくために，抽出した情報を新たに追加学習していく必要がある．しかし，プライバシー情報と抽出した情報が必ずしも正しいとは限らない．そこで，現在処理中の文章に対するスコア加算の際に以前の文章から解析し抽出したプライバシー情報のスコアを反映させる．これにより，1 つの文章だけではプライバシー情報としてスコアが低い語句でも何度も抽出すると，プライバシー情報が漏洩する危険性が高いことを示せる．

そこで，文章が複数に及ぶ場合での抽出した情報の蓄積手法，曖昧性解消のための忘却といった追加学習の手法を検討する．

5.2. 実現方式

5.1.で示した知識の組み合わせにより抽出できる情報の確かさを比較し，重要度によりそれぞれの知識に付加されるべきスコアを定める．

また，そのスコアを加算することで，複数文章にまたがる場合の知識の蓄積手法を確立する．

5.3. スコア割り当ての方法

(1) 検知した結果に対するスコア付加

図３に日記例を示す．各日記において情報漏洩の大小の違いを考え，それぞれの知識の重要度を検討する．ここでは，住所に関するプライバシー情報の知識として，【NG ワード】に「調布」，【種類語】に「住む」，【特徴語】に「町」が登録されているとする．各日記例から検知できる語句をまとめたものを表２に示す．

日記１からは，【種類語】とそれに対応する

日記1	調布に住んでいます。
日記2	小島町に住んでいます。
日記3	仙川に住んでいます。
日記4	マンションに住んでいます。
日記5	小島町のスーパーで買い物をした。
日記6	仙川のスーパーで買い物をした。

図３ 日記例

表2 各日記から抽出される語句

	NGワード	特徴語	種類語	抽出されるべきプライバシー情報	スコア
日記1	調布		住む	調布	6pt
日記2		町	住む	小島町	3pt
日記3			住む	仙川	2pt
日記4			住む	—	—
日記5		町		小島町	1pt
日記6				仙川	0pt

【NG ワード】が抽出される．そのため，明らかに危険であると判断される．

日記２では，【種類語】とそれに対応する【特徴語】が抽出される．そのため，「小島町」がプライバシー情報であると推測できる．

日記３では，【種類語】のみのためプライバシー情報は抽出されない．しかし，本来は「仙川」を地名と判断しプライバシー情報と推測するべきである．

日記４では，【種類語】のみのためプライバシー情報は抽出されない．日記５は，【特徴語】のみの抽出であり，「小島町」は地名であるがプライバシー情報であるかは判断できない．日記６は【特徴語】も【種類語】も含まれていないため検知できない．

以上から，３つの知識の重要度を比較すると，【NG ワード】は絶対にプライバシー情報である．日記３のように【種類語】のみが含まれる文章の場合でもプライバシー情報と推測されるべき場合がある．さらに，日記５のように【特徴語】のみが含まれる文章では，プライバシー情報と推測しきれないことがある．

よって，３種類の知識は【NG ワード】，【種類語】，【特徴語】の順に重要度が低くなることが分かる．これより，スコアの絶対値を決定する．本来は多数の例文から妥当な結果が出るように調整する必要があるが，今回はこれらの例文を参考に暫定的ではあるが【NG ワード】４pt，【種類語】２pt，【特徴語】１pt とする．

(2) 複数文章におけるスコア加算

複数文章にまたがる解析を行うために，文章から抽出されたプライバシー情報をスコアとともに学習していく．現在処理中の文章に過去に学習したプライバシー情報が含まれる場合は，検知したプライバシー情報に付加されているスコアも現在の処理に加算する．

たとえば，日記２と日記５が順に投稿された場合を考える．日記２から「小島町」を住所に関するプライバシー情報と推測し，知識として学習する．このとき【種類語】と【特徴語】が含まれるため(１)で定めたスコアから「小島町」は 3pt であるとして学習する．そして，日記５が投稿されると，「小島町」をすでに学習しているため，プライバシー情報として検知する．日記５における「小島町」のスコアは【特徴語】の１pt であるが，日記２よりすでに 3pt が付加されているため，加算し４pt になる．このように，スコアの増加により，情報の確かさを示すことが可能になり，複数文章にまたがったプライバシー情報の検知ができる．なお，複数の文章にまたがるスコア計算方法について，今回は単純加算としたが，より適切な計算方法を今後検討する．

6. 結論

自然言語の日記からプライバシー情報を自動検知するためには，システムが膨大な知識を準備する必要がある．また，複数の文章にまたがった検知が必要である．検知に必要な知識を分析し，全てのユーザに共通の知識とユーザ固有の知識に分類できることを示した．さらに，格フレーム検索ツールを用いて５億文のテキストコーパスから，共通知識を自動抽出する方法を提案した．また，プライバシー情報を表す語句である【NG ワード】や【種類語】，【特徴語】が文章中にどれだけ含まれるかによって検知したプライバシー情報にスコアをつける方法を提案し，複数文章にわたってスコアを加算することで，複数文章からの情報検知を可能とした．

今後の課題を以下に示す．

(１)提案方式の実装・評価

(２)ユーザ固有の知識の準備手法の検討

(３)複数文章からのスコア加算方法の改良

参考文献

[1] 片岡春乃， 他; 意味と面白さを維持する自然言語情報の開示制御技術の提案—SNS のプライバシー保護への試適用—， 情報処理学会 第 38 回コンピュータセキュリティ研究会 (2007)

[2] 片岡春乃， 他; 自然言語情報の開示制御技術 DCNL の実現に向けて—プライバシー情報検知手法—， 情報処理学会 第 40 回コンピュータセキュリティ研究会 (2008)

[3] 渡辺夏樹， 他; 自然言語文からのプライバシー情報検知システム—自然言語情報の開示制御技術 DCNL の実現（１）—， 情報処理学会 第 44 回コンピュータセキュリティ研究会 (2009)

[4] 京都大学黒橋研究室．“格フレーム検索”．http://reed.kuee.kyoto-u.ac.jp/cf-search/，(最終閲覧 2009/01/29)