特別研究報告

題目

# ソフトウェア開発ドキュメント匿名化ツールの試作とその評価

指導教員
楠本 真二 教授

報告者
江川　翔太

平成 26 年 2 月 14 日

大阪大学 基礎工学部 情報科学科

平成 25 年度 特別研究報告

# ソフトウェア開発ドキュメント匿名化ツールの試作とその評価

江川　翔太

## 内容梗概

ソフトウェア工学研究では，実際のプロジェクトで開発されたドキュメント，ソースコード等のソフトウェア開発ドキュメントの調査や，得られた研究成果を実際のプロジェクトに適用・評価することが重要である．一方，ソフトウェア開発ドキュメントの多くは著作権や機密保持の点で公開することは難しい．さらに，個別の共同研究においても，企業から大学へ成果物の提供を行う際に機密情報の匿名化を行う必要がある．しかし，複数のドキュメントを人手で匿名化しなければならず，その作業にかかる人的コストが研究での実プロジェクトソフトウェア開発ドキュメントの利用を困難にしている．

本研究では，匿名化にかかるコストを削減しソフトウェア開発ドキュメントの研究利用の促進を目的として，ソフトウェア開発ドキュメントの匿名化を支援するツールを開発する．ツールでは対象ドキュメントに対する形態素解析・類語認識や利用者とのインタラクティブなやりとりを通じて，固有表現やコンテキスト情報（固有表現以外に間接的に固有表現を特定できるもの）の匿名化を支援する．匿名化ツールを実装し，実プロジェクトの成果物を対象に匿名化がどの程度効果的に行えるかどうかを評価した．その結果，ツールが自動で行う匿名化の場合，適合率が 0.827，再現率が 0.788 となった．また，ユーザによる短時間の補正で，完全な匿名化を実現できた．

## 主な用語

ソフトウェア開発ドキュメント
匿名化
固有表現
コンテキスト情報
形態素解析

目 次

# 表 目 次

# 図 目 次

## 1　まえがき

産学連携研究とは企業と大学が共同で行う研究である．産学連携研究を行うことで科学論文の生産性が向上することが確認されている [1]．ソフトウェア工学は高品質で低コストなソフトウェアの開発補助を目的とする分野で，コードクローン検出技術 [2] やソフトウェアの見積技術 [3] のような技術を扱う．ソフトウェアの開発・保守は主に企業が行うため，ソフトウェア工学では実プロジェクトのソースコードや要求定義書，設計書のようなソフトウェア開発ドキュメントの利用が重要である．ソフトウェアの開発・保守は主に企業が行う．そのため，ソフトウェア工学では実プロジェクトのソースコードや要求定義書，設計書のようなソフトウェア開発ドキュメントの利用が重要である．しかし，企業から提供されるドキュメントは開発担当者の名前や連絡先のような個人情報や，顧客との間に守秘義務がある情報を含むため，企業から大学へソフトウェア開発ドキュメントの提供を行う際は匿名化する必要がある．一方，ドキュメントに含まれる情報は膨大で人手による匿名化はコストがかかる．よって，大学での実プログラムのソフトウェア開発ドキュメント利用は困難で，ツールによるソフトウェア開発ドキュメントの匿名化コスト削減が求められる．医療関係書類を対象とした自動匿名化手法が提案されている [4, 5, 6, 7, 8, 9, 10] が，医療関係書類で匿名化すべき情報とソフトウェア開発ドキュメントで匿名化すべき情報は異なる．

そこで，本研究ではソフトウェア開発ドキュメントの匿名化を支援するツールを開発する．本ツールは，ドキュメント中に出現する人名，地名，組織名といった固有表現を抽出し，別の記号に置換して匿名化を行う．このとき，正しく抽出されなかった固有表現はユーザの補助で修正する．また，参照することで匿名化した固有表現を推測可能な表現もユーザの補助で匿名化する．ツールを用いることで匿名化にかかるコストが減少し，ソフトウェア開発ドキュメントの研究利用が促進されると考えられる．

匿名化ツールを実装し，ある実プロジェクトの成果物を対象に匿名化がどの程度効果的に行えるかどうかを評価した．その結果，ツールが自動で行う匿名化の場合，適合率が 0.827，再現率が 0.788 となった．また，ユーザによる短時間の補正で，完全な匿名化を実現できた．

以降，2 章では研究の背景となる諸用語と関連研究について述べる．3 章では匿名化ツールの概要と匿名化手法の詳細について述べる．4 章では作成したツールの評価実験と結果についての考察を述べる．最後に 5 章で本報告のまとめを述べる．

## 2　準備

本章では研究の背景となる諸用語と関連研究について簡単に述べる．

### 2.1　匿名化

本研究において，匿名化とは連絡先情報の置換，固有表現の置換およびコンテキスト情報の置換を指す．次小節以降で連絡先情報の置換，固有表現の置換およびコンテキスト情報の置換の詳細を述べる．

#### 2.1.1　連絡先情報の置換

連絡先情報は，メールアドレス，電話番号，郵便番号，住所の丁番地および URL を指す．ドキュメント中に出現するこれらの情報を，記号で置換する事で匿名化を行う．例えば，メールアドレスの置換例を図 1 に示す．

#### 2.1.2　固有表現の置換

本研究において，固有表現とは IREX[11] によって規定された 8 種類の分類の内，人名，地名および組織名を指す．ただし，人名では苗字と名前を区別する．ドキュメント中に出現するこれらの文字列を，記号で置換する事で匿名化を行う．例えば，人名と組織名の置換例を図 2 に示す．

また，ドキュメント中に同一の固有表現が複数出現する場合，それらを同一の記号で置換する．例えば，図 3 では「豊中」という地名が 2 度出現するが，それらを「地名 1」という同一の記号で置換を行っている．

#### 2.1.3　コンテキスト情報の置換

ある文字列をもとに固有表現を推測できる場合，そのような文字列をコンテキスト情報と定義し匿名化の対象とする．例として，銀行名，銀行の支店数および従業員数を含む銀行向け勘定系システム開発ドキュメントの匿名化を考える．固有表現の匿名化では銀行名の匿名化は成されるが，支店数と従業員数の匿名化は成されない．．ここで，ドキュメントの読み手が意図的に各銀行の支店数と従業員数を調査すると銀行名が特定されてしまう恐れがある．この場合，コンテキスト情報は支店数や従業員数となる．これを防ぐためコンテキスト情報は匿名化する必要がある．


図 1: 連絡先情報の置換

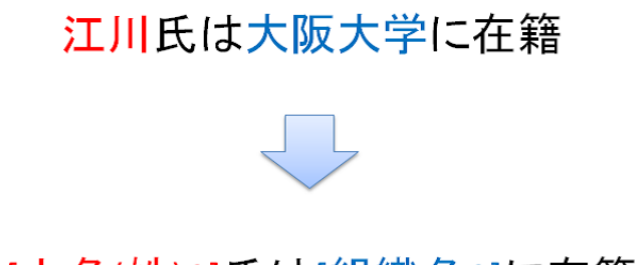


図 2: 固有表現の置換

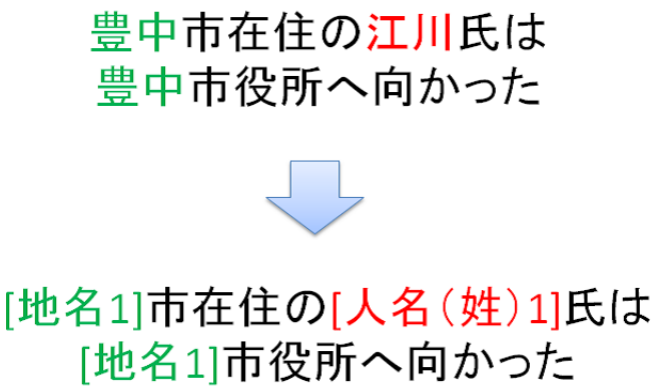


図 3: 同一固有表現の置換

| | |
|---|---|
| 大阪大学 | 名詞,固有名詞,組織,大阪大学,オオサカダイガク |
| の | 助詞,連体化,の,ノ |
| 教務 | 名詞,一般,教務,キョウム |
| システム | 名詞,一般,システム,システム |
| を | 助詞,格助詞,一般,を,ヲ |
| 開発 | 名詞,サ変接続,開発,カイハツ |
| する | 動詞,自立,サ変・スル,基本形,する,スル |

図 4: 形態素解析の例

### 2.2 関連研究

#### 2.2.1 形態素解析

テキスト解析技術として形態素解析が研究されている [12, 13, 14, 15]．形態素解析とは，自然言語で記述された文章を意味を持つ最小単位に分割しそれぞれの品詞を判別する技術である．形態素解析ツールを用いることでドキュメント内に出現する固有表現の大半を判別することが可能となる．例として,「大阪大学の教務システムを開発する」という文章を形態素解析し品詞に分解した結果を図 4 に示す.

#### 2.2.2 自動匿名化

ドキュメントの自動的な匿名化については，カルテや退院時要約等の医療関係書類を自動的に匿名化する手法が提案されている [4, 5, 6, 7, 8, 9, 10]．医療関係書類に関しては，HIPAA[16] により書類を他の研究機関に提供する際に匿名化する必要のある情報が明確に定められている．また，匿名化の対象となるファイルはテキストファイルである．これに対しソフトウェア開発ドキュメントは主なファイル形式が Word・Excel であり，匿名化する必要のある情報は明確に定められておらず，コンテキスト情報の存在もあり医療関係書類の自動匿名化手法を適用することは難しい．つまり，ソフトウェア開発ドキュメントを対象とした匿名化手法や実際に匿名化を行うのに適したツールは存在しない.

## 2.3 ドキュメントの匿名化における問題点

### 2.3.1 コンテキスト情報の選定

大学を対象とした履修登録システム開発ドキュメントの場合，コンテキスト情報は学部名や総学生数等となる．一方，銀行を対象とした勘定系システム開発ドキュメントの場合，コンテキスト情報は支店数や従業員数となる．このようにコンテキスト情報はドキュメントの内容によって異なるため自動的な匿名化は困難である．

### 2.3.2 固有表現の分類

固有表現は同一の表記で複数のものを指す場合がある．例えば「大津」という人名には同名の地名が存在する．また，一般名詞と判別がつかない固有表現も存在する．例えば,「森」は一般名詞であるが，文脈により人名を指す場合がある．

### 2.3.3 表記ゆれ

表記ゆれとは同じ意味の語句について異なる文字表記が付されることである．人名では漢字表記と読み仮名表記，組織名では略称や別称といった表記ゆれが存在する．例えば，大阪大学という組織には阪大や OsakaUniversity 等の異なる表記が存在する．同一ドキュメント中に出現するこれらの表現を異なる記号で置換した場合，単一の組織を指しているにも関わらず，読み手に複数の組織として認識される恐れがある．そのため異なる表記を同一の記号で置換する必要がある．

組織名の略称，別称に関しては Web 上の類語辞典を参照し同一の記号で置換する．

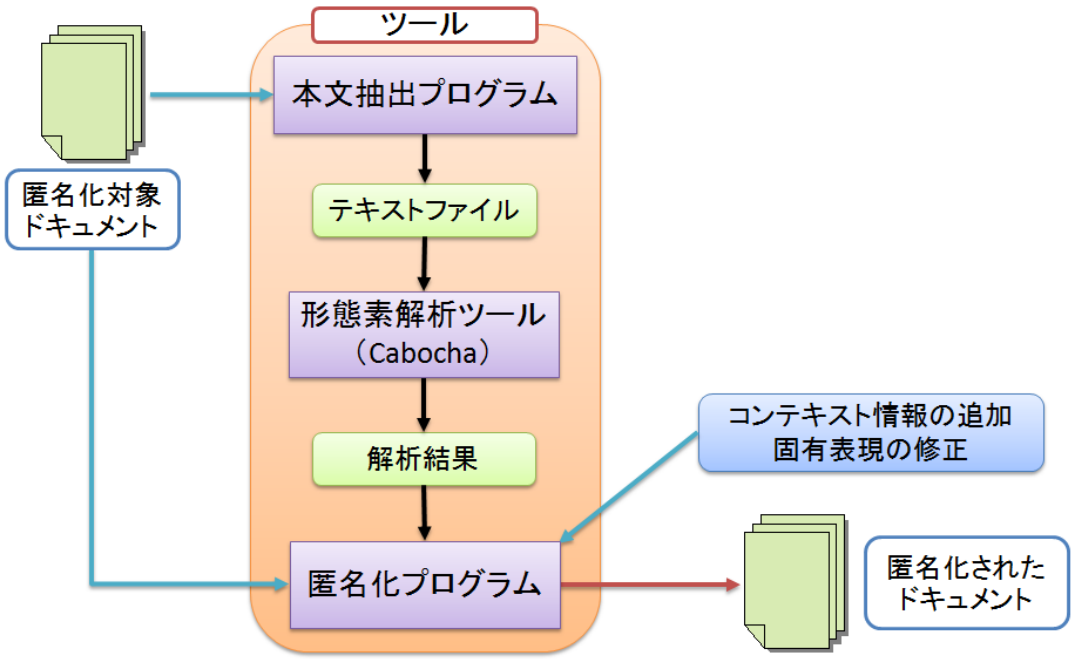


図 5: ツールの動作概要

## 3　提案する匿名化手法

本章では，ソフトウェア開発ドキュメントの匿名化手法について説明する．

### 3.1　概要

図 5 にツールの動作概要を示す．ツールは匿名化対象ドキュメントを入力とし，ドキュメント内に出現する連絡先情報，固有表現およびコンテキスト情報の置換を行う．出力として，匿名化されたドキュメントを出力する．ツールは本文抽出プログラム，形態素解析ツールおよび匿名化プログラムから構成される．以降，各部分の動作について説明する．

### 3.2　本文抽出プログラム

匿名化の対象となるドキュメント群から本文を抽出しテキストファイルに変換する．

### 3.3　形態素解析ツール

テキストファイルに記述された文章を，品詞情報を付加して分解する．

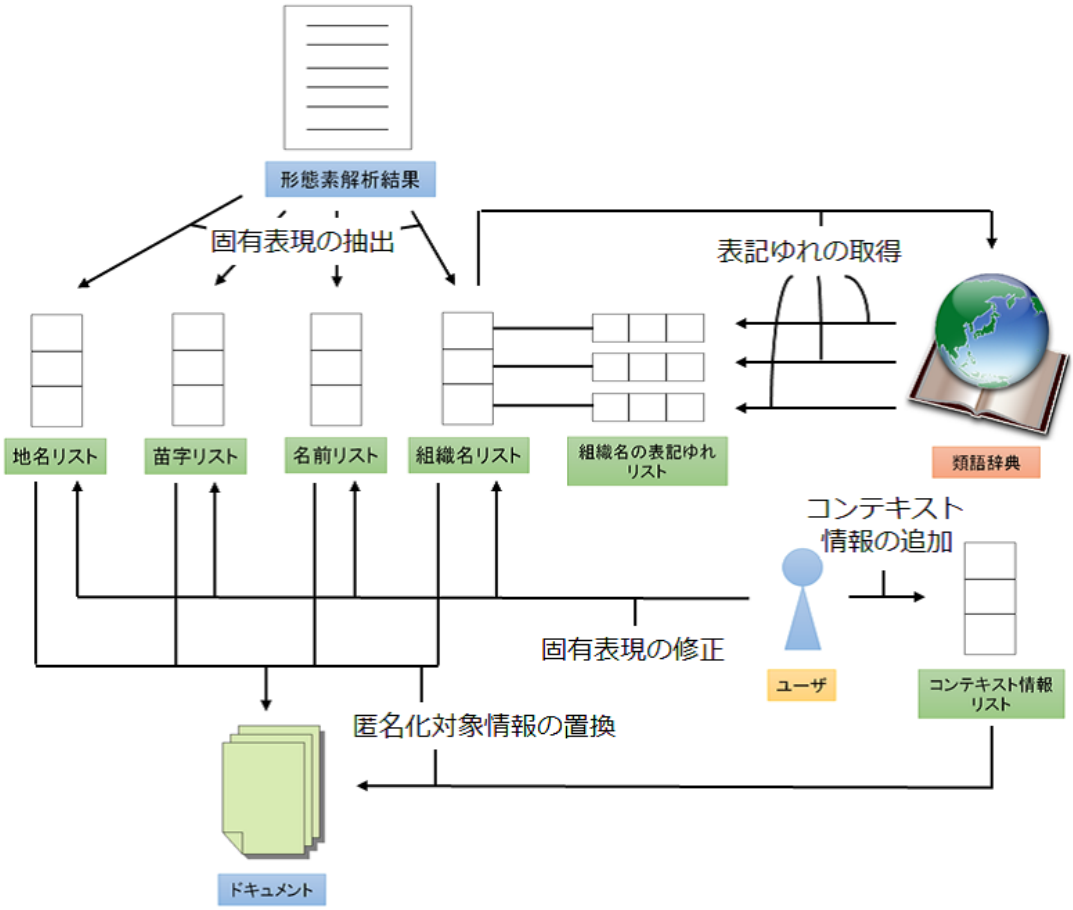


図 6: 匿名化プログラムの動作の流れ

### 3.4 匿名化プログラム

ドキュメント内に出現する固有表現，連絡先情報およびコンテキスト情報を置換し匿名化する．匿名化プログラムは，固有表現の抽出，固有表現の修正，表記ゆれの取得，コンテキスト情報の取得および匿名化対象情報の置換という 5 つの動作を行う．匿名化プログラムの動作の流れを図 6 に示す．

以降で各動作の詳細について説明する．

#### 3.4.1 固有表現の抽出

形態素解析結果が記述されたテキストファイルを参照し，固有表現として判別された文字列を抽出する．この時，固有表現を人名，地名および組織名に分類する．

#### 3.4.2 固有表現の修正

抽出された固有表現をユーザに提示し，分類を誤った固有表現の再分類，抽出されなかった固有表現の追加および誤って固有表現として抽出された文字列の削除を依頼する．

#### 3.4.3 表記ゆれの取得

組織名をクエリとして類語辞典を検索し表記ゆれを取得する．

#### 3.4.4 コンテキスト情報の取得

ユーザに入力を依頼しコンテキスト情報として扱う文字列を取得する．

#### 3.4.5 匿名化対象情報の置換

匿名化対象ドキュメントから本文を抽出し連絡先情報，固有表現およびコンテキスト情報を置換する．

### 3.5 実装

本節では，ツールの各部分の実装について述べる．ツールは Java[17] を用いて実装した．現在のところツールが対象としているファイル形式は，txt，doc および xls である．

#### 3.5.1 本文抽出プログラム

Apache POI ライブラリ [18] を用いて Word および Excel 形式のファイル本文を抽出する．そして，本文をテキストファイルへ出力する．

### 3.5.2 形態素解析ツール

形態素解析ツールとして Cabocha[13] を用いる．Cabocha は外部プログラムであるため，ProcessBuilder を用いて実行する．このとき，Cabocha への入力として本文出力プログラムが出力したテキストファイルを与える．Cabocha は形態素解析結果が記述されたテキストファイルを出力する．

### 3.5.3 匿名化プログラム

固有表現の抽出では，まず，固有表現を分類するために ArrayList を用いて，苗字リスト，名前リスト，地名リストおよび組織名リストを作成する．次に，形態素解析ツールの解析結果を 1 行ずつ読み込み「固有名詞」と記述された行の先頭の文字列を固有表現として抽出する．そして，図 4 で示されるような，同一行に記述された固有表現の分類を参照し各リストに記憶する．

固有表現の修正では，まず，抽出した固有表現をリストごとにユーザに提示する．次に，それぞれのリストについて固有表現の追加と削除をユーザへ依頼する．

表記ゆれの取得では，類語辞典として Weblio 類語辞典 [19] を用いる．Weblio 類語辞典の検索結果の URL は以下の様になっている．

`http://thesaurus.weblio.jp/content/`検索ワードの `URL` エンコード結果

まず，リストに保存した組織名と同数の表記ゆれリストを作成する．次に，組織名の URL エンコードを行い検索結果ページの URL を取得する．そして，取得した URL を用いて検索結果ページにアクセスし HTML を取得する．最後に，表記ゆれが記述された部分を HTML から抽出し表記ゆれをリストに記憶する．

コンテキスト情報の取得では，まず，コンテキスト情報を記憶するためのリストを作成する．次に，コンテキスト情報の入力をユーザへ依頼する．そして，ユーザが入力した文字列をリストへ追加する．

連絡先情報の置換では，まず，ドキュメントから本文を 1 行ずつ抽出する．そして，正規表現によるパターンマッチを行い，マッチした場合は連絡先情報を別の記号に置換する．連絡先情報の正規表現と置換後の文字列を表 1 に示す．

固有表現の置換では，まず，ドキュメントから本文を 1 行ずつ抽出する．そして，各固有表現リストを参照し，抽出した文に固有表現が含まれている場合は別の文字列に置換する．各リストに含まれる固有表現の置換後の文字列を表 2 に示す．x はリストに記憶された順番を表す．

コンテキスト情報の置換では，まず，ドキュメントから本文を 1 行ずつ抽出する．そして，

コンテキスト情報リストを参照し，抽出した文にコンテキスト情報が含まれている場合は「その他 x」という文字列に置換する．x はリストに記憶された順番を表す．

### 3.6 適用例

本節では図 7 に示すツールの適用対象となるドキュメントの一部を用いて匿名化を説明する．ドキュメントのファイル形式は xls である．

本文抽出プログラムは，図 7 に示すドキュメントから本文を抽出しテキストファイルに変換する．変換した結果を図 8 に示す．

形態素解析ツールは，図 reffigL に示すテキストファイルに対し形態素解析を行う．形態素解析結果の一部を図 9 に示す．

図 9 で示した解析結果から固有表現を抽出すると,「江川」は苗字,「吹田」,「和歌山」および「栄谷」は地名,「和歌山大学」は組織名として分類される．しかし，本来人名として扱われるべき「吹田」という固有表現が地名として分類されている．そのため，ユーザに固有表現の修正を依頼し,「吹田」を人名として再分類する．

固有表現を修正した後，Weblio 類語辞典にアクセスし，組織名の表記ゆれを取得する．Weblio 類語辞典で「和歌山大学」を検索した結果を図 10 に示す．

和歌山大学の略称および別称として,「Wakayama University」,「和大」,「和歌山大学付属学校」および「わかやまだいがく」が提示されるため，これらの文字列を表記ゆれとして取得する．その後，ユーザに入力を依頼し，コンテキスト情報を取得する．そして，連絡先情報，固有表現およびコンテキスト情報の置換を行うことでドキュメントを匿名化する．匿名

表 1: 連絡先情報の正規表現と置換後の文字列

| 連絡先情報 | 正規表現 | 置換後の文字列 |
|---|---|---|
| メールアドレス | [\w\.\-]+@(?:[\w\-]+\.)+[\w\-]+ | xxxx@xxxx |
| 電話番号 | \d{1,4}?-\d{1,4}?-\d{1,4} | xxxx-xxxx-xxxx |
| 郵便番号 | \d{3}-\d{4} | xxx-xxxx |
| 住所の丁番地 | ((丁目\|番地?\|号)?[-ー－]?)$ | xxx-xxx-xxx |
| URL | (https?\|ftp)(:\/\/[-_.! *\'()a-zA-Z0-9;\/?:\@\&=+$,%#]+) | URL |

表 2: 連絡先情報の正規表現と置換後の文字列

| リスト名 | 置換後の文字列 |
|---|---|
| 組織名リスト | 組織名 x |
| 苗字リスト | 人名（姓）x |
| 名前リスト | 人名（名）x |
| 地名リスト | 地名 x |

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | 要求仕様書 | | | | | |
| 2 | | | | | | |
| 3 | 1. はじめに | | | | | |
| 4 | | | | | | |
| 5 | **顧客** | 和歌山大学 | | | | |
| 6 | **窓口** | システム工学部　吹田教授 | | | | |
| 7 | **所属** | 〒640-8510　和歌山県和歌山市栄谷930 | | | | |
| 8 | | TEL/FAX 073-123-4567 携帯 090-1234-5678 | | | | |
| 9 | | MAIL suita@wakayama-u.ac.jp | | | | |
| 10 | | | | | | |
| 11 | | | | | | |
| 12 | **本書の目的** | | | | | |
| 13 | | 本書は和歌山大学(以下和大)教務関係業務システムの作成において必要となる要求仕様をまとめたものです。 | | | | |
| 14 | | | | | | |
| 15 | | | | | | |
| 16 | **作成者** | 江川 | | | | |
| 17 | | | | | | |
| 18 | | | | | | |

図 7: 匿名化対象ドキュメントの一部

```
要求仕様書

1. はじめに

顧客　　和歌山大学
窓口　システム工学部　吹田教授
所属　〒640-8510　和歌山県和歌山市栄谷９３０
  TEL/FAX 073-123-4567 携帯 090-1234-5678
  MAIL suita@wakayama-u.ac.jp

本書の目的
  本書は和歌山大学（以下和大）教務関係業務システムの作成において必要となる要求仕様をまとめたものです。

作成者　　江川
```

図 8: テキストファイルへの変換結果

```
要求        名詞,サ変接続,要求,ヨウキュウ
仕様        名詞,一般,仕様,シヨウ
書          名詞,接尾,一般,書,ショ
1           名詞,数
.           名詞,サ変接続
はじめ      名詞,副詞可能,はじめ,ハジメ
に          助詞,格助詞,一般,に,ニ
顧客        名詞,一般,顧客,コキャク
            記号,空白
和歌山大学  名詞,固有名詞,組織,和歌山大学,ワカヤマダイガク
窓口        名詞,一般,窓口,マドグチ
システム    名詞,一般,システム,システム
工学部      名詞,一般,工学部,コウガクブ
            記号,空白
吹田        名詞,固有名詞,地域,一般,吹田,スイタ
教授        名詞,一般,教授,キョウジュ
所属        名詞,サ変接続,所属,ショゾク
〒          記号,一般,〒,ユウビンバンゴウ
640         名詞,数
-           名詞,サ変接続
8510        名詞,数
            記号,空白
和歌山      名詞,固有名詞,地域,一般,和歌山,ワカヤマ
県          名詞,接尾,地域,県,ケン
和歌山      名詞,固有名詞,地域,一般,和歌山,ワカヤマ
市          名詞,接尾,地域,市,シ
栄谷        名詞,固有名詞,地域,一般,栄谷,サカエダニ
```

図 9: 形態素解析結果（一部）

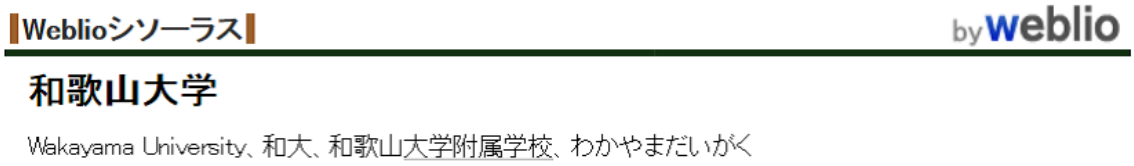


図 10: weblio シソーラスの検索結果

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | 要求仕様書 | | | | | |
| 2 | | | | | | |
| 3 | 1. はじめに | | | | | |
| 4 | | | | | | |
| 5 | 顧客 | 組織名1 大学 | | | | |
| 6 | 窓口 | その他1 学部　人名(姓)2教授 | | | | |
| 7 | 所属 | 〒xxx-xxxx　地名1県地名1市地名2xxxxx | | | | |
| 8 | | TEL/FAX xxxx-xxxx-xxxx 携帯 xxx-xxxx-xxxx | | | | |
| 9 | | MAIL xxxx@xxxx | | | | |
| 10 | | | | | | |
| 11 | | | | | | |
| 12 | 本書の目的 | | | | | |
| 13 | | 本書は組織名1大学(以下組織名1)教務関係業務システムの作成において必要となる要求仕様をまとめたものです。 | | | | |
| 14 | | | | | | |
| 15 | | | | | | |
| 16 | 作成者 | 人名(姓)1 | | | | |
| 17 | | | | | | |
| 18 | | | | | | |

図 11: 匿名化されたドキュメントの一部

化されたドキュメントの一部を図 11 に示す.

# 4　評価実験

本章では評価実験の結果及び考察について述べる．

## 4.1　実験対象

ツールの適用対象とするドキュメントに IT Spiral 実プロジェクト教材 [20] を用いる．これは，和歌山大学の履修登録システムの開発の際に得られた要求定義書やプロジェクト管理書，画面設計書等のドキュメントが教材として大学に提供されたものである．本実験では，和歌山大学を対象としたシステム開発であることが推測可能となる，学部名やコース名をコンテキスト情報して扱う．ドキュメントが全 97 ファイル存在し，ファイルサイズは合計 16.9MB，ドキュメント内に出現する固有表現数は 170，コンテキスト情報数は 21 となる．

## 4.2　実験内容

ツールの有用性を調査するために 2 つの実験を行った．評価基準としてそれぞれの実験において適合率と再現率を定義する．

### 4.2.1　実験 1

ツールがどの程度自動で匿名化を行うかを調査するために，ユーザ入力を受け付けずにツールをドキュメントに適用した．適用後，ツールが匿名化した文字列数と正確に匿名化した固有表現数をカウントし適合率と再現率を求めた．

実験 1 における適合率はツールがどれだけ正確に匿名化を行ったかを示す指標であり，以下のように定義する．

$$\text{適合率} = \frac{\text{ツールが正確に匿名化した固有表現数}}{\text{ツールが匿名化した文字列数}}$$

実験 1 における再現率はツールがどれだけ情報を漏らさずに匿名化を行ったかを示す指標であり，以下のように定義する．

$$\text{再現率} = \frac{\text{ツールが正確に匿名化した固有表現数}}{\text{ドキュメント内に出現する固有表現数}}$$

### 4.2.2　実験 2

ツールが正確に匿名化を行うことが可能かどうかを調査するためにツールをドキュメントに適用した．適用後，ツールが匿名化した文字列数と正確に匿名化した固有表現数とコンテキスト情報数をカウントし適合率と再現率を求めた．

実験 2 における適合率はツールがどれだけ正確に匿名化を行ったかを示す指標であり，以下のように定義する．

$$\text{適合率} = \frac{\text{ツールが正確に匿名化した固有表現数} + \text{コンテキスト情報数}}{\text{ツールが匿名化した文字列数}}$$

実験 2 における再現率はツールがどれだけ情報を漏らさずに匿名化を行ったかを示す指標であり，以下のように定義する．

$$\text{再現率} = \frac{\text{ツールが正確に匿名化した固有表現数} + \text{コンテキスト情報数}}{\text{ドキュメント内に出現する固有表現数} + \text{コンテキスト情報数}}$$

### 4.3 実験結果

#### 4.3.1 実験 1

実験 1 の結果を表 3 に示す．表の縦軸はそれぞれ以下のことを表している．

**固有表現数** ドキュメント中に出現する固有表現数

**ツール出力** ツールが匿名化した文字列数

**出力正解数** ツールが正確に匿名化した固有表現数

**適合率** ツールが匿名化した文字列の内，匿名化すべき固有表現数の割合

**再現率** ドキュメント内に出現する固有表現数の内，ツールが匿名化した固有表現数の割合

#### 4.3.2 実験 2

実験 2 の結果を表 4 に示す．表の縦軸はそれぞれ以下のことを表している．

**情報数** ドキュメント中に出現する固有表現数+コンテキスト情報数

**ツール出力** ツールが匿名化した文字列数

**出力正解数** ツールが正確に匿名化した情報数

表 3: 実験 1 の結果

| 固有表現数 | ツール出力 | 出力正解数 | 適合率 | 再現率 |
|---|---|---|---|---|
| 170 | 162 | 134 | 0.827 | 0.788 |

**適合率** ツールが匿名化した文字列の内，匿名化すべき情報数の割合

**再現率** ドキュメント内に出現する情報数の内，ツールが匿名化した情報数の割合

また，ユーザによる入力が行われた時間は 10 分程度となった．

### 4.4 考察

#### 4.4.1 実験 1

実験 1 においてツールが行った匿名化の内，正確に匿名化が行われなかった例について述べる．

実験 1 において適合率は 0.827 となっている．適合率はツールが誤った匿名化を行った場合に低下する．ツールが誤って匿名化した文字列として，大学の授業で用いられる教科書の著者名と「大津」という人名が挙げられる．著者名の匿名化を行った場合，著者が開発社もしくは顧客に関係する人物であるという誤解を招く恐れがあるため，匿名化する必要はない.「大津」という人名に関しては，同表記の地名が存在するため文脈により誤って地名と分類される場合が存在した．

また，再現率は 0.788 となっている．再現率はツールが匿名化すべき固有表現を匿名化できなかった場合に低下する．ツールが匿名化できなかった固有表現として，人名の読み仮名と「日本システム技術株式会社」という一般名詞が組み合わさった組織名が挙げられる．人名の読み仮名が固有表現として抽出されないのは，複数の一般名詞が組み合わさった文字列として扱われるためである．

以上のような誤った匿名化が行われないようにするためには，辞書をあらかじめ用意する必要がある．例えば，開発に関わる人物や組織名のリストを作成しツールがそれを参照すれば実験 1 における誤りは全て正確に匿名化されると考えられる．

また，長期に渡って同開発社のプロジェクトのドキュメントを匿名化する予定である場合，頻出する組織名や人名を学習することで，誤った匿名化を防ぐことが可能になると考えられる．

表 4: 実験 2 の結果

| 情報数 | ツール出力 | 出力正解数 | 適合率 | 再現率 |
|---|---|---|---|---|
| 191 | 191 | 191 | 1.000 | 1.000 |

#### 4.4.2 実験 2

実験 2 において，ユーザに固有表現の修正とコンテキスト情報の追加を 10 分程度依頼することで適合率と再現率は共に 1.000 となった．この結果によりツールの有用性が確認できた．

今後の課題として，現在ユーザ入力に依存しているコンテキスト情報取得の自動化を行うことが挙げられる．また，実際の開発現場で作成されたソフトウェア開発ドキュメントに対してツールを適用し，ユーザ入力が介在する場合でも匿名化に失敗するケースがあるかを調査することが考えられる．

## 5　あとがき

本研究では，ソフトウェア開発ドキュメントの匿名化を支援するツールを開発した．

また，ツールの有用性を評価するため，複数のソフトウェア開発ドキュメントを対象にツールを適用し，匿名化の適合率と再現率を調査した．その結果，ユーザ入力が介在しない匿名化の適合率は 0.827，再現率は 0.788 となり，ユーザが 10 分程度修正を行うことで適合率，再現率は共に 1.000 となった．

本研究の今後の課題は以下の通りである．

- 対応ファイル形式の拡張や，匿名化すべき情報がまとめられた辞書の作成の支援といった，ツールの機能追加を行う．
- 実際の現場で作成されたソフトウェア開発ドキュメントにツールを適用し，有用性を調査する．

## 謝辞

本研究を行うにあたり，理解あるご指導を賜り，常に励まして頂きました 楠本 真二 教授に心より感謝申し上げます.

本研究の全過程を通し，終始熱心かつ丁寧なご指導を頂きました 岡野 浩三 准教授に深く感謝申し上げます.

本研究に関して，的確なご助言ご指導を頂きました 井垣 宏 特任准教授に心より感謝申し上げます.

本研究を行うにあたり，日常の議論の中でご助言を頂きました 肥後 芳樹 助教に心より感謝申し上げます.

本報告を行うにあたりご指導，ご協力を頂き，さらに日常でも声をかけて頂きました大阪大学大学院情報科学研究科コンピュータサイエンス専攻博士前期課程 2 年の 佐々木 幸広 氏，同 榛葉　浩章 氏に深く感謝申し上げます.

また研究室生活の中で相談に乗って頂き，また励まして頂きました 大阪大学大学院情報科学研究科コンピュータサイエンス専攻博士前期課程 1 年の大田 崇史 氏，同 楠 野明 氏，同 藤田 悠矢 氏に心より感謝申し上げます.

その他の楠本研究室の皆様のご協力に心より感謝致します.

最後に，本研究に至るまでに，講義，演習，実験等でお世話になりました大阪大学基礎工学部情報科学科の諸先生方に，この場を借りて心から御礼申し上げます.

## 参考文献


[1] 七丈直弘, 馬場靖憲. 産学連携が大学の科学研究に与える影響の定量分析 (産学官連携 (1)). 研究・技術計画学会, 2006.

[2] 肥後芳樹, 楠本真二, 井上克郎. コードクローン検出とその関連技術. 電子情報通信学会論文誌 D, Vol. 91, No. 6, pp. 1465–1481, 2008.

[3] 松川文一, 楠本真二, 井上克郎, 英繁雄, 前川祐介. ユースケースポイント計測支援ツールの実装とその適用. 情報処理学会研究報告, Vol. 144, No. 30, pp. 91–98, 2004.

[4] Stephane Meystre, F Friedlin, Brett South, Shuying Shen, and Matthew Samore. Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC medical research methodology*, Vol. 10, No. 1, p. 70, 2010.

[5] Ishna Neamatullah, Margaret M Douglass, H Lehman Li-wei, Andrew Reisner, Mauricio Villarroel, William J Long, Peter Szolovits, George B Moody, Roger G Mark, and Gari D Clifford. Automated de-identification of free-text medical records. *BMC medical informatics and decision making*, Vol. 8, No. 1, p. 32, 2008.

[6] Kostas Pantazos, Soren Lauesen, and Soren Lippert. De-identifying an EHR database-Anonymity, Correctness and Readability of the Medical Record. *Proceedings of MIE2011*, 2011.

[7] György Szarvas, Richárd Farkas, and Róbert Busa-Fekete. State-of-the-art anonymization of medical records using an iterative machine learning framework. *Journal of the American Medical Informatics Association*, Vol. 14, No. 5, pp. 574–580, 2007.

[8] Karen Tu, Julie Klein-Geltink, Tezeta F Mitiku, Chiriac Mihai, and Joel Martin. De-identification of primary care electronic medical records free-text data in Ontario, Canada. *BMC medical informatics and decision making*, Vol. 10, No. 1, p. 35, 2010.

[9] Özlem Uzuner, Yuan Luo, and Peter Szolovits. Evaluating the state-of-the-art in automatic de-identification. *Journal of the American Medical Informatics Association*, Vol. 14, No. 5, pp. 550–563, 2007.

[10] Özlem Uzuner, Tawanda C Sibanda, Yuan Luo, and Peter Szolovits. A de-identifier for medical discharge summaries. *Artificial intelligence in medicine*, Vol. 42, No. 1, pp. 13–35, 2008.

[11] Satoshi Sekine and Hitoshi Isahara. IREX: IR & IE Evaluation Project in Japanese. In *LREC*, pp. 1977–1980, 2000.

[12] Taku Kudo. Mecab: Yet another part-of-speech and morphological analyzer. *http://mecab. sourceforge. net/*, 2005.

[13] K Taku. Cabocha: Yet another japanese dependency structure analyzer. Technical report, Technical report, Nara Institute of Science and Technology, 2004.

[14] 黒橋禎夫. 日本語形態素解析システム juman version 3.5. *(http://www-lab25. kuee. kyoto-u. ac. jp/nl-resource/juman. html)*, 1998.

[15] 松本裕治, 北内啓, 山下達雄, 平野善隆, 松田寛, 浅原正幸. 日本語形態素解析システム「茶筌」 version 2.0 使用説明書. *NAIST-IS-TR99012*, 1999.

[16] Centers for Disease Control, Prevention (CDC, et al. HIPAA privacy rule and public health. Guidance from CDC and the US Department of Health and Human Services. *MMWR. Morbidity and mortality weekly report*, Vol. 52, p. 1, 2003.

[17] Oracle technology network for java developers — oracle technology network — oracle. *http://www.oracle.com/technetwork/java/index.html*.

[18] Apache poi - the java api for microsoft documents. *http://poi.apache.org/*.

[19] 類語辞典・シソーラス - weblio 辞書. *http://thesaurus.weblio.jp/*.

[20] 実プロジェクト教材. *http://it-spiral.ist.osaka-u.ac.jp/project/education.html*.