

# 単語分散表現と単語N-gramを用いた セキュリティ要件分類手法の改良

オープンソースソフトウェア工学研究室

宮崎智己

# 研究の背景

- 開発者は利用者の要件に沿ったシステムを開発する必要がある
  - 業務内容や利用環境によって利用者が求める品質特性は異なる

## ISO25010[1]における品質特性

機能適合性  
性能効率性  
互換性

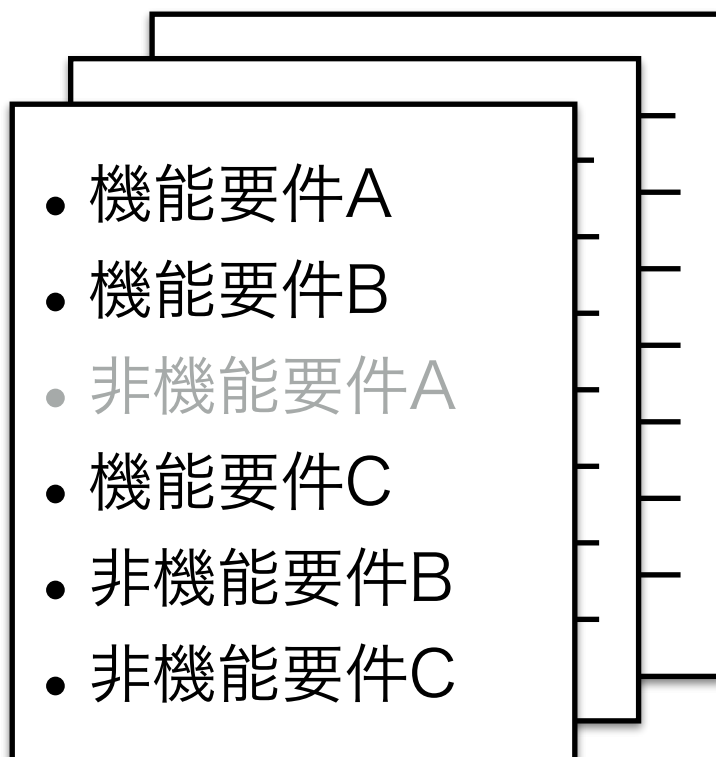
信頼性  
セキュリティ  
保守性

使用性  
移植性

- 品質特性を満たすように機能・非機能要件を定義

# 機能・非機能要件の特定

- 開発者はシステム構築に必要な機能・非機能要件を要件定義書から特定し理解する必要がある
  - 開発者は手作業で確認



要件定義書



機能要件A  
機能要件B  
機能要件C  
非機能要件B  
非機能要件C

# 非機能要件の見落とし

---

- 非機能要件は機能要件と比べて見落とされやすい
  - 非機能要件の多くは機能要件に含まれる[2]
    - ▶ ユーザ認証機能の要件（満たすとセキュリティが向上）
  - 機能要件と比べて曖昧になりやすく判断が困難[3]
    - ▶ ”すぐに”, ”不満がでない程度に”, など

非機能要件の見落としは  
開発工程の手戻りやシステム障害の原因となってしまう

[2]Eckhardt, J., Vogelsang, A. and Fernández, D. M.: Are "Non-functional" Requirements Really Non-functional?: An Investigation of Non-functional Requirements in Practice, Proc. of the 38th International Conference on Software Engineering (ICSE '16), pp. 832-842 (2016).

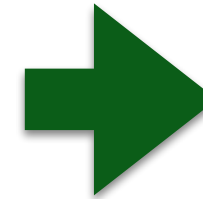
[3]Borg, A., Yong, A., Carlshamre, P. and Sandahl, K.: The Bad Conscience of Requirements Engineering: An Investigation in Real-World Treatment of Non-Functional Requirements, Proc. of the 3rd Conference on Software Engineering Research and Practice in Sweden (SERPS '03), pp. 1-8 (2003).

# 非機能要件分類手法[4,5]

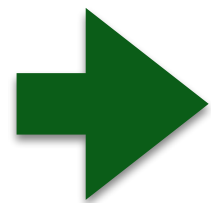
- ナイーブベイズやSVMを用いた非機能要件分類手法が提案されている

学習データ

|     |                                 |     |
|-----|---------------------------------|-----|
| (1) | Allows role-based access to EH  | 機密性 |
| (2) | Provides timely access to infor | 機密性 |
| (3) | Data used during user authent   | 真正性 |
| (4) | D2D performs identity manage    | 真正性 |

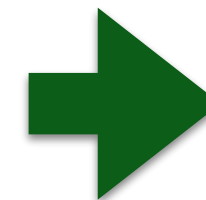


- 記号・ストップワードの除去
- 単語の正規化



|     | access | role | data | • • | perform |
|-----|--------|------|------|-----|---------|
| (1) | 1      | 1    | 0    | • • | 0       |
| (2) | 0      | 0    | 0    | • • | 0       |
| (3) | 0      | 0    | 1    | • • | 0       |
| (4) | 0      | 0    | 0    | • • | 1       |

1gram BOWの作成



ナイーブベイズ  
SVM

[4]Wen Zhang, Ye Yang, Qing Wang, Fengdi Shu, An Empirical Study on Classification of Non-Functional Requirements, Proceedings of the 23rd International Conference on Software Engineering & Knowledge Engineering (SEKE'2011)

[5]John Slankas and Laurie Williams, Automated Extraction of Non-functional Requirements in Available Documentation, 1st International Workshop on Natural Language Analysis in Software Engineering (NaturaLiSE), (2013)

# 既存手法の問題点

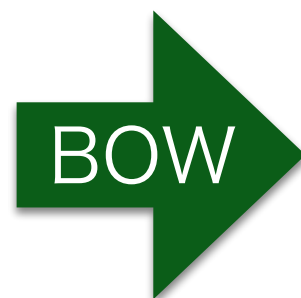
1. 単語の類似性を考慮しておらず類義語に対応できない

- “**login**”と“**logon**”を違う単語として判断してしまう

2. 単語の順序関係を考慮しておらず複合語に対応することができない

ア.**Log in** as oscar or mac2002 or 1117.

イ.All POS systems connected to the EHRi must record **in**  
an audit **log**.



|   | in | log | POS | ・ ・ | 1117 |
|---|----|-----|-----|-----|------|
| ア | 1  | 1   | 0   | ・ ・ | 1    |
| イ | 1  | 1   | 1   | ・ ・ | 0    |

# 研究の目的・アプローチ

---

- 目的

- 非機能要件自動分類手法の精度向上
  - 本研究ではセキュリティ要件の詳細な分類に着目

- アプローチ

- 類義語の統一
  - シソーラスを用いた方法
  - 分散表現手法を用いた方法
- 単語の順序関係を特徴量に用いる
  - 単語N-gramを用いた特徴ベクトルの作成

# 提案手法の概要

---

1. 前処理
2. シソーラスを用いた類義語の統一
3. 分散表現を用いた類義語の統一
4. N-gramを特徴量に用いた特徴ベクトルの作成
5. 分類器の構築
  - ナイーブベイズ, SVMを用いる

# 前処理

---

- ピリオドやカンマといった記号の除去
- 文字を大文字から小文字に変換
- 単語に分割
  - 冠詞を除去
  - レマタイズ処理
    - 単語を基本形に変換する処理：（例）複数形を単数形に変換

All POS systems connected to the EHRI must record in an audit log.



'all', 'po', 'system', 'connected', 'to', 'ehri', 'must', 'record', 'in',  
'audit', 'log'

# シソーラスを用いた類義語の統一

- SEThesaurus[6]：ソフトウェア開発に関する単語の表記ゆれを解消するためのシソーラス
  - スタックオーバーフローの投稿に対してWord2Vecとレーベンシュタイン距離を適用して類疑義を抽出
    - 分散表現の類似度が高く，文字列の類似度が高い単語
    - 手作業で調整済み

SEThesaurus

• javascript  
• javascript  
• js

The system need be developed with **js.**

統一

The system is developed in **Javascript.**

# 分散表現手法を用いた類義語の統一

---

- 分散表現：単語を任意の次元のベクトルで表現
- 以下の3つの分散表現手法を適用
  - Word2Vec[7]
  - Fasttext[8]
  - SO\_Word2Vec[9]
    - スタックオーバーフローを学習したWord2Vec学習済みモデル

[7]Mikolov, T., Sutskever, I., Chen, K., Corrado, G. and Dean, J.: Distributed Representations of Words and Phrases and Their Compositionality, Proceedings of the 26th International Conference on Neural Information Processing Systems (NIPS '13), USA, Curran Associates Inc., pp. 3111–3119 (2013).

[8]P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, “Enriching word vectors with subword information,” Transactions of the Association for Computational Linguistics, vol.5, pp.135–146, (2017).

[9]V. Efstathiou, C. Chatzilenas, and D. Spinellis, “Word embeddings for the software engineering domain,” Proc. of the 15th International Conference on Mining Software Repositories (MSR '18), pp.38–41, (2018)

# Word2Vec

- ニューラルネットワークを用いた単語分散表現手法
- Skip-gram（中心単語から周囲に出現する単語を予測）を適用

$$\frac{1}{T} \sum_{t=1}^T \sum_{-c \leq j \leq c, j \neq 0} \log p(w_{t+j} | w_t)$$

最大になるよう  
重みを学習

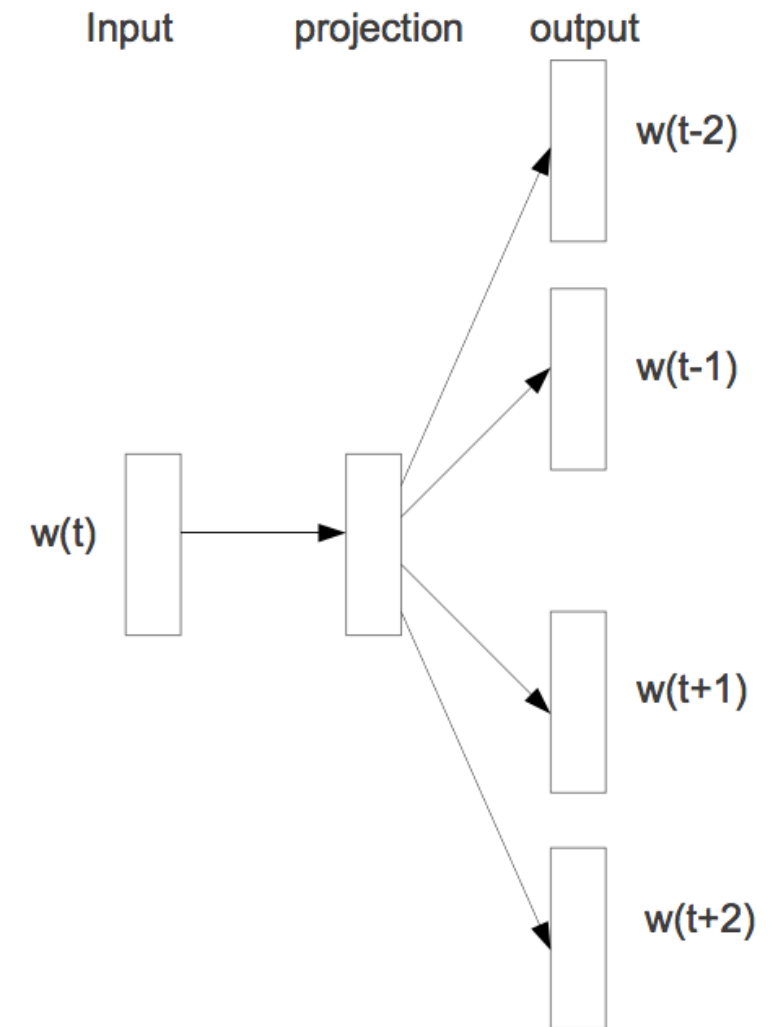
$$p(w_{t+j} | w_t) = \frac{\exp(v_{w_{t+j}}^T v_{w_t})}{\sum_{w=1}^W \exp(v_w^T v_{w_t})}$$

$w_t$ : 単語

$T$ : 学習する単語の総数

$C$ : 予測する周辺単語の数

$p(w_{t+j} | w_t)$ :  $w_t$ の時に $w_{t+j}$ が発生する確率



Skip-gramの概要図

# Fasttext

.....

- Word2Vecの学習に単語の文字列を考慮したモデル
  - 入力単語を文字Nグラムに分割
  - Nグラムをそれぞれベクトル化し，ベクトルの重みを足し合わせる

例) "where"を入力データとした際，N=3の時

where→<wh, whe, her, ere, re>

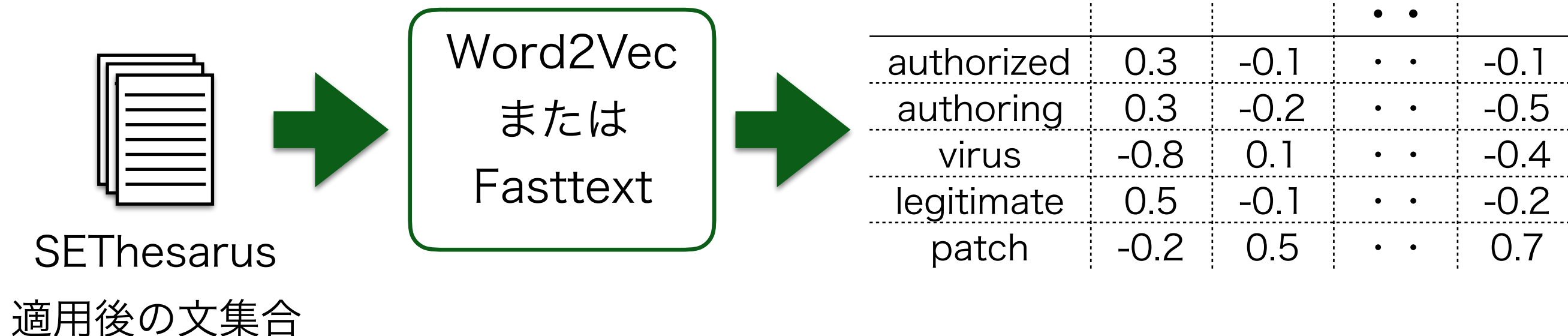
$$p(w_{t+j}|w_t) = \frac{\exp(v_{w_{t+j}}'^T v_{w_t})}{\sum_{w=1}^W \exp(v_w'^T v_{w_t})} \quad \rightarrow \quad p(w_{t+j}|w_t) = \frac{\exp(\sum_{g=1}^G z_{v_{w_t},g}^T v_{w_{t+j}})}{\sum_{w=1}^W \exp(v_w'^T v_{w_t})}$$

$z_{v_{w_t},g}^T$  : 単語  $w_t$  の  $g$  番目のNグラムのベクトル

$G$  : 単語をNグラムに分割してできた文字列の総数

# 分散表現手法を用いた類義語の統一

- 前処理後の要件文に対して分散表現手法を適用し分散表現を取得（Word2Vec, Fasttextのみ）
  - 出現回数が少ない単語は除外
  - 計算コスト削減のためにネガティブサンプリングを実施
    - 学習の際にランダムに偽の入力をし，正解単語の出現確率を低下させる方法



.....

類義語とみよしく抽出

# 分散表現手法を用いた類義語の統一

---

- 単語クラスターの構成単語の一つに単語を統一

- overridden by **authorized** users
- **authoring** HTML documents
- impersonating a **legitimate** user

単語クラスター

authorized  
authoring  
legitimate

統一

- overridden by **authorized** users
- **authorized** HTML documents
- impersonating a **authorized** user

# 単語の順序関係を特徴量にする方法

## 1. N-gramのみを特徴量とした特徴ベクトルの作成

- (例) 2-gram

|     | log in | in as | • | in audit | audit log |
|-----|--------|-------|---|----------|-----------|
| (1) | 1      | 1     | • | 0        | 0         |
| (2) | 0      | 0     | • | 0        | 1         |

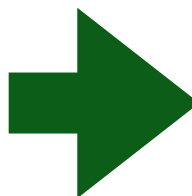
## 2. 1-gramとN-gramを特徴量とした特徴ベクトルの作成

|     | log | in | log in | in as | log in as | • • | audit log |
|-----|-----|----|--------|-------|-----------|-----|-----------|
| (1) | 1   | 1  | 1      | 1     | 1         | • • | 0         |
| (2) | 1   | 1  | 0      | 0     | 0         | • • | 1         |

# 1-gramとN-gramを特徴量とした特徴ベクトルの作成

- Multiword Expression Distance (MED) [10]を適用して特徴量にするN-gramを抽出
  - MEDの値に外れ値除去を行い残ったN-gramを特徴量として抽出

$$MED(g) = \frac{df(w_1, \dots, w_n)}{df(g)} \quad \begin{array}{l} df(w_1, \dots, w_n): \text{Nグラム}g\text{を構成する単語} \\ \text{を全て含む文書数} \\ df(g): \text{Nグラム}g\text{を含む文書数} \end{array}$$

|                                                                                                                            |                                                                                       |                                                                                   |
|----------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------|
| <ul style="list-style-type: none"><li>• apple orange banana</li><li>• apple orange banana</li><li>• orange apple</li></ul> |  | $MED(apple\ orange) = \frac{2}{3}$ $NED(apple\ orange\ banana) = \frac{2}{2} = 1$ |
|----------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------|

# 評価実験

- 以下の組み合わせを比較

|   | SEThesaurus | 分散表現手法      |   | 特徴ベクトル          |
|---|-------------|-------------|---|-----------------|
| 1 | あり          | Word2Vec    | × | 1-gram          |
| 2 | あり          | Fasttext    |   | 2-gram          |
| 3 | あり          | SO_Word2vec |   | 3-gram          |
| 4 | なし          | SO_Word2vec |   | 1-gram + N-gram |

- Word2Vec・Fasttextのパラメータを変えて比較

- ▶ 次元数：200（固定）
- ▶ 最小出現回数：5, 10, 15
- ▶ ウィンドウサイズ（学習に使う周辺単語数）：5, 10, 15
- ▶ ネガティブサンプリング：5, 10, 15

- 評価指標：Precision・Recall・F値

# データセット

.....

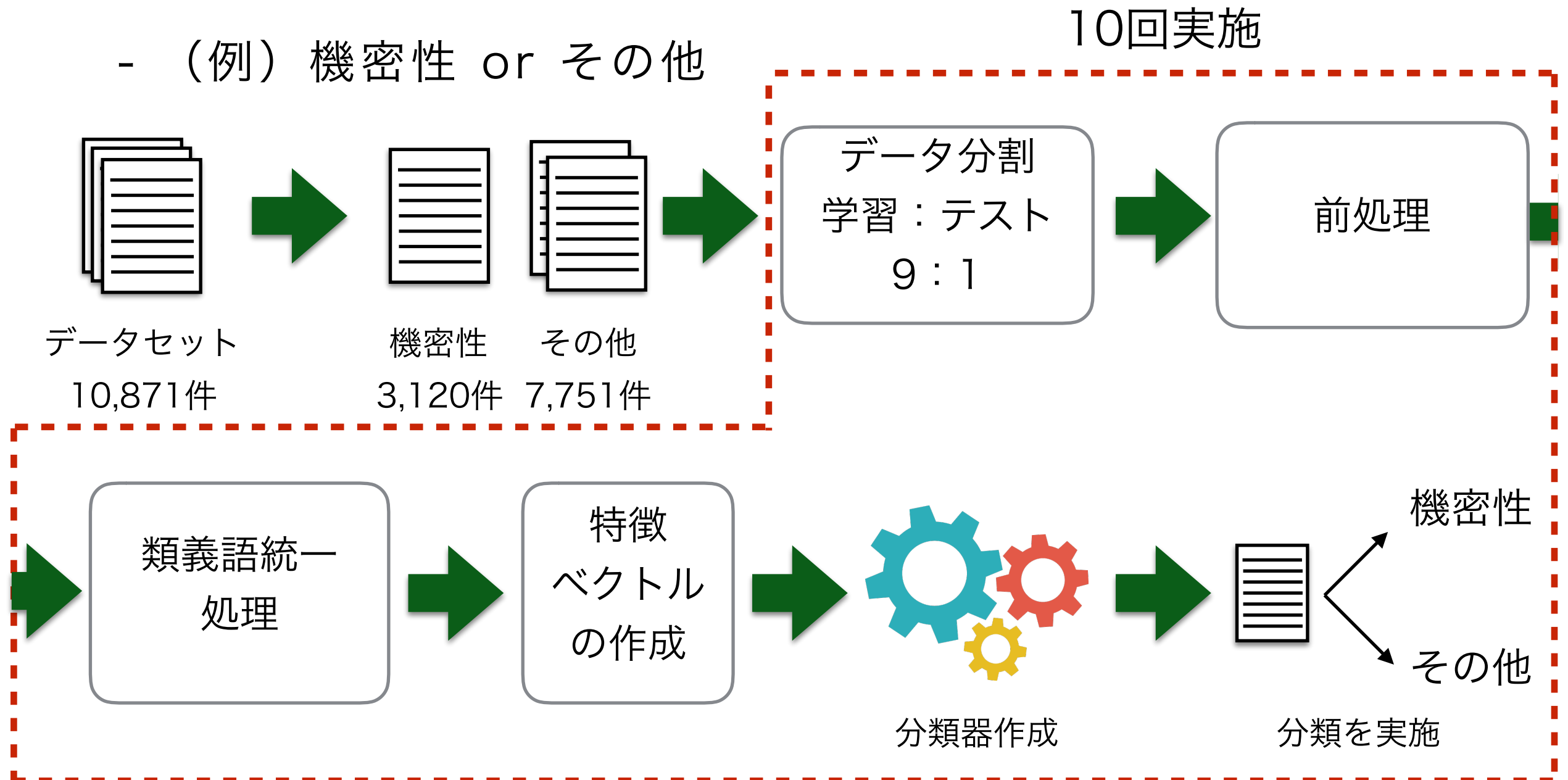
- 5つの医療系システムの要件定義書・機能要求
  - 1つの文に複数のセキュリティ要件が含まれる場合あり
  - データ数が一定以上ある機密性，完全性，責任追跡性を対象

| クラス          | 件数     |
|--------------|--------|
| 機密性          | 3,120  |
| 完全性          | 3,386  |
| 可用性          | 174    |
| 責任追跡性        | 3,882  |
| 真正性          | 311    |
| プライバシー       | 246    |
| Not Security | 6,617  |
| total        | 10,871 |

# 実験手順

- 各セキュリティ要件ごとに2値分類器を作成し10分割  
交差検証をそれぞれ実施

- (例) 機密性 or その他



# ナイーブベイズ・機密性の実験結果

| 特徴ベクトル | SE Thesaurus | 分散  | 最小 | ウィンドウ | ネガティブ | 閾値 | Precision    | Recall       | F値           |
|--------|--------------|-----|----|-------|-------|----|--------------|--------------|--------------|
| 既存手法   |              |     |    |       |       |    | 0.712        | 0.786        | 0.747        |
| 2-gram | あり           | SO  | 5  | 5     | 5     | 8  | <b>0.754</b> | <b>0.811</b> | <b>0.781</b> |
| 2-gram | あり           | SO  | 5  | 5     | 5     | 10 | 0.754        | 0.811        | 0.781        |
| 2-gram | あり           | SO  | 5  | 5     | 5     | 12 | 0.753        | 0.812        | 0.781        |
| 2-gram | あり           | W2V | 10 | 15    | 15    | 2  | 0.754        | 0.811        | 0.781        |
| 2-gram | あり           | Ft  | 15 | 15    | 15    | 2  | 0.753        | 0.811        | 0.781        |

上位5パターンを表示

# ナイーブベイズ・完全性の実験結果

| 特徴ベクトル | SE Thesaurus | 分散  | 最小 | ウィンドウ | ネガティブ | 閾値 | Precision    | Recall       | F値           |
|--------|--------------|-----|----|-------|-------|----|--------------|--------------|--------------|
| 既存手法   |              |     |    |       |       |    | 0.708        | 0.805        | 0.753        |
| 2-gram | あり           | W2V | 15 | 10    | 5     | 2  | <b>0.741</b> | <b>0.830</b> | <b>0.783</b> |
| 2-gram | あり           | Ft  | 5  | 10    | 15    | 2  | 0.742        | 0.828        | 0.783        |
| 2-gram | あり           | W2V | 5  | 15    | 5     | 2  | 0.745        | 0.825        | 0.782        |
| 2-gram | あり           | W2V | 15 | 10    | 15    | 2  | 0.741        | 0.829        | 0.782        |
| 2-gram | あり           | W2V | 10 | 15    | 10    | 2  | 0.740        | 0.830        | 0.782        |

上位5パターンを表示

# ナイーブベイズ・責任追跡性の実験結果

| 特徴ベクトル | SE Thesaurus | 分散 | 最小 | ウィンドウ | ネガティブ | 閾値 | Precision | Recall | F値    |
|--------|--------------|----|----|-------|-------|----|-----------|--------|-------|
| 既存手法   |              |    |    |       |       |    | 0.769     | 0.856  | 0.810 |
| 2-gram | あり           | Ft | 10 | 15    | 15    | 2  | 0.837     | 0.836  | 0.836 |
| 2-gram | あり           | Ft | 10 | 10    | 15    | 2  | 0.836     | 0.836  | 0.836 |
| 2-gram | あり           | Ft | 5  | 15    | 10    | 2  | 0.838     | 0.835  | 0.836 |
| 2-gram | あり           | Ft | 15 | 10    | 5     | 2  | 0.838     | 0.834  | 0.836 |
| 2-gram | あり           | Ft | 10 | 15    | 5     | 2  | 0.835     | 0.837  | 0.836 |

上位5パターンを表示

# SVM・機密性の実験結果

| 特徴ベクトル | SE Thesaurus | 分散  | 最小 | ウィンドウ | ネガティブ | 閾値 | Precision    | Recall       | F値           |
|--------|--------------|-----|----|-------|-------|----|--------------|--------------|--------------|
| 既存手法   |              |     |    |       |       |    | 0.779        | 0.756        | 0.768        |
| N-gram | あり           | Ft  | 5  | 15    | 5     | 2  | <b>0.812</b> | <b>0.777</b> | <b>0.794</b> |
| N-gram | あり           | W2V | 10 | 10    | 15    | 2  | 0.813        | 0.776        | 0.793        |
| N-gram | あり           | W2V | 10 | 10    | 10    | 2  | 0.808        | 0.779        | 0.793        |
| N-gram | あり           | Ft  | 15 | 15    | 15    | 2  | 0.813        | 0.773        | 0.792        |
| N-gram | あり           | W2V | 10 | 10    | 5     | 2  | 0.812        | 0.773        | 0.792        |

上位5パターンを表示

# SVM・完全性の実験結果

| 特徴ベクトル | SE Thesaurus | 分散  | 最小 | ウィンドウ | ネガティブ | 閾値 | Precision    | Recall       | F値           |
|--------|--------------|-----|----|-------|-------|----|--------------|--------------|--------------|
| 既存手法   |              |     |    |       |       |    | 0.781        | 0.767        | 0.773        |
| N-gram | あり           | W2V | 15 | 15    | 5     | 2  | <b>0.805</b> | <b>0.780</b> | <b>0.792</b> |
| N-gram | なし           | SO  | 5  | 5     | 5     | 10 | 0.806        | 0.779        | 0.792        |
| N-gram | なし           | SO  | 5  | 5     | 5     | 16 | 0.804        | 0.781        | 0.792        |
| N-gram | なし           | なし  | 0  | 0     | 0     | 0  | 0.806        | 0.779        | 0.792        |
| N-gram | なし           | SO  | 5  | 5     | 5     | 2  | 0.806        | 0.778        | 0.791        |

上位5パターンを表示

# SVM・責任追跡性の実験結果

| 特徴ベクトル | SE Thesaurus | 分散 | 最小 | ウィンドウ | ネガティブ | 閾値 | Precision    | Recall       | F値           |
|--------|--------------|----|----|-------|-------|----|--------------|--------------|--------------|
| 既存手法   |              |    |    |       |       |    | 0.834        | 0.813        | 0.823        |
| N-gram | なし           | SO | 5  | 5     | 5     | 8  | <b>0.858</b> | <b>0.831</b> | <b>0.844</b> |
| N-gram | なし           | SO | 5  | 5     | 5     | 16 | 0.859        | 0.830        | 0.844        |
| N-gram | なし           | SO | 5  | 5     | 5     | 4  | 0.858        | 0.831        | 0.844        |
| N-gram | なし           | SO | 5  | 5     | 5     | 10 | 0.857        | 0.831        | 0.844        |
| N-gram | なし           | SO | 5  | 5     | 5     | 18 | 0.857        | 0.831        | 0.844        |

上位5パターンを表示

# 実験のまとめ

---

- N-gram処理
  - ナイーブベイズは2-gram, SVMはN-gramの場合が精度が良い
- SEThesaurus
  - ナイーブベイズではありの場合の方が良い
- 分散表現
  - 各セキュリティ要件ごとにパラメータやアルゴリズムが異なる

# TODO

---

- 残りの3種類（可用性・真正性・プライバシー）の要件の実験
  - Word2Vec, SO\_Word2Vecは終了
  - fasttextは残り約2/3
- 実験結果の考察
  - セキュリティ要件ごとに分散表現に違いがあった理由を調査
  - ナイーブベイズとSVMの違いも影響している？
- 修論執筆