# 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

- 業務内容や利用環境によって利用者がシステムに求める品質特性は異なる
  - 自動改札システムでは性能効率性が，金融系のシステムではセキュリティが求められる

ISO25010[1]における品質特性

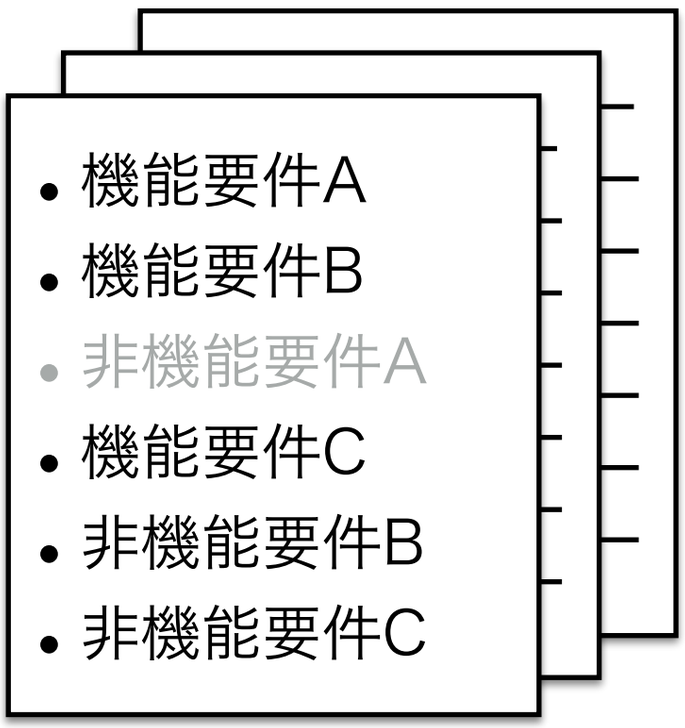
| | | |
|---|---|---|
| 機能適合性 | 信頼性 | 使用性 |
| 性能効率性 | セキュリティ | 移植性 |
| 互換性 | 保守性 | |

- 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

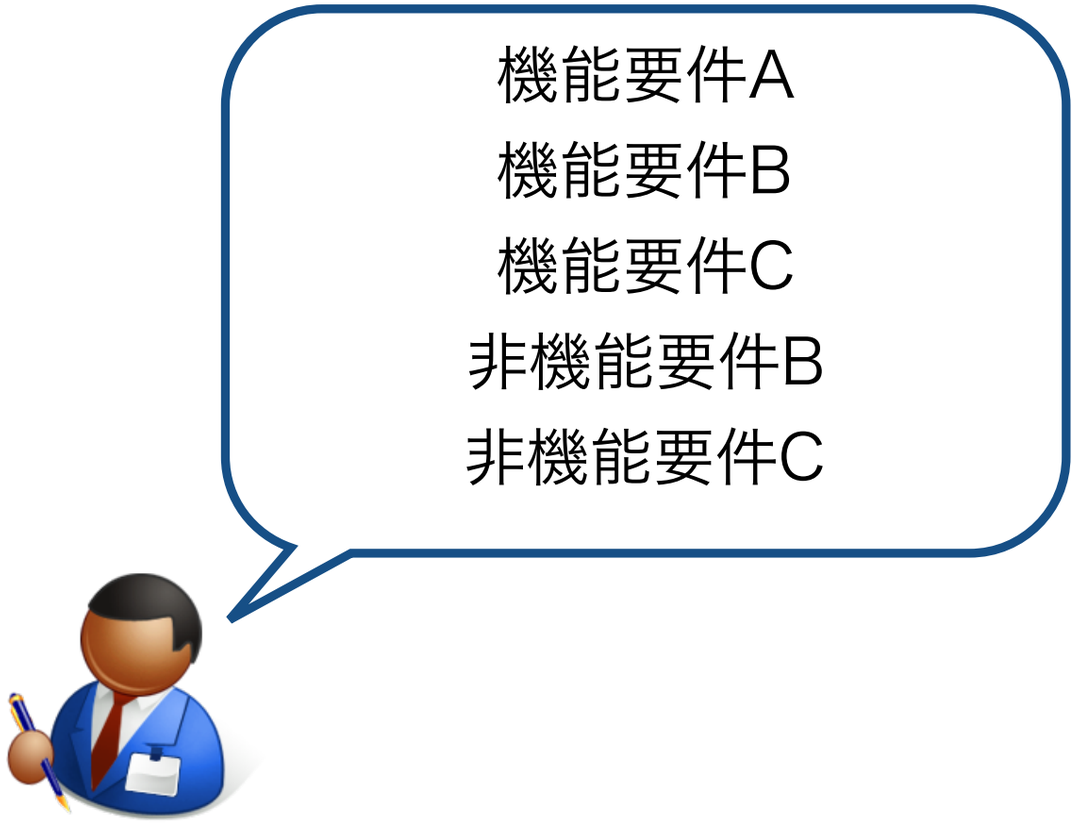
[1]ISO/IEC: ISO/IEC 25010 - Systems and software en- gineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - System and software quality models, Technical report (2010).

# 機能・非機能要件の特定

- 開発者は要件定義書からシステム構築に必要な機能・非機能要件を手作業で特定する必要がある




要件定義書



# 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]
    - 認証機能に関する要件（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 Fern´andez, 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 In- vestigation 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]

- ナイーブベイズ，SMO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

SMO

[4]M. Riaz, J. King, J. Slankas, and L. Williams, “Hidden in plain sight: Automatically identifying security requirements from natural language artifacts,” Proc. of 22nd International Requirements Engineering Conference (RE ’14), pp.183–192, Aug. 2014.

# 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 |

全く同じと判断

# 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

前処理

↓

シソーラスを用いた
類義語の統一

↓

分散表現を用いた
類義語の統一

↓

N-gramを特徴量とした
特徴ベクトルの作成

↓

分類器の構築

# 提案手法の概要

前処理

↓

シソーラスを用いた
類義語の統一

↓

分散表現を用いた
類義語の統一

↓

N-gramを特徴量とした
特徴ベクトルの作成

↓

分類器の構築

# 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'

# 提案手法の概要

前処理

↓

シソーラスを用いた
類義語の統一

↓

分散表現を用いた
類義語の統一

↓

N-gramを特徴量とした
特徴ベクトルの作成

↓

分類器の構築

# 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
  - Stack Overflowの投稿に対してWord2Vecとレーベンシュタイン距離を適用し類義語を抽出
    - 分散表現の類似度が高く，文字列の類似度が高い単語

SEThesaurus

- javascript
- javasript
- js

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

The system is developed in **Javascript.**

[6]Chen, C., Xing, Z. and Wang, X.: Unsupervised Software-specific Morphological Forms Inference from In- formal Discussions, Proc. of the 39th International Con- ference on Software Engineering (ICSE '17), Piscat- away, NJ, USA, IEEE Press, pp. 450–461 (2017).

# 提案手法の概要

前処理

↓

シソーラスを用いた
類義語の統一

↓

**分散表現を用いた**
**類義語の統一**

↓

N-gramを特徴量とした
特徴ベクトルの作成

↓

分類器の構築

# 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する手法

- 以下の3つの分散表現手法を適用
  - Word2Vec[7]
  - Fasttext[8]
  - SO_Word2Vec[9]
    - Stack Overflow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 Asso- ciates 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 do- main," Proc. of the 15th International Conference on Mining Software Repositories (MSR '18), pp.38–41, (2018)

# Word2Vec

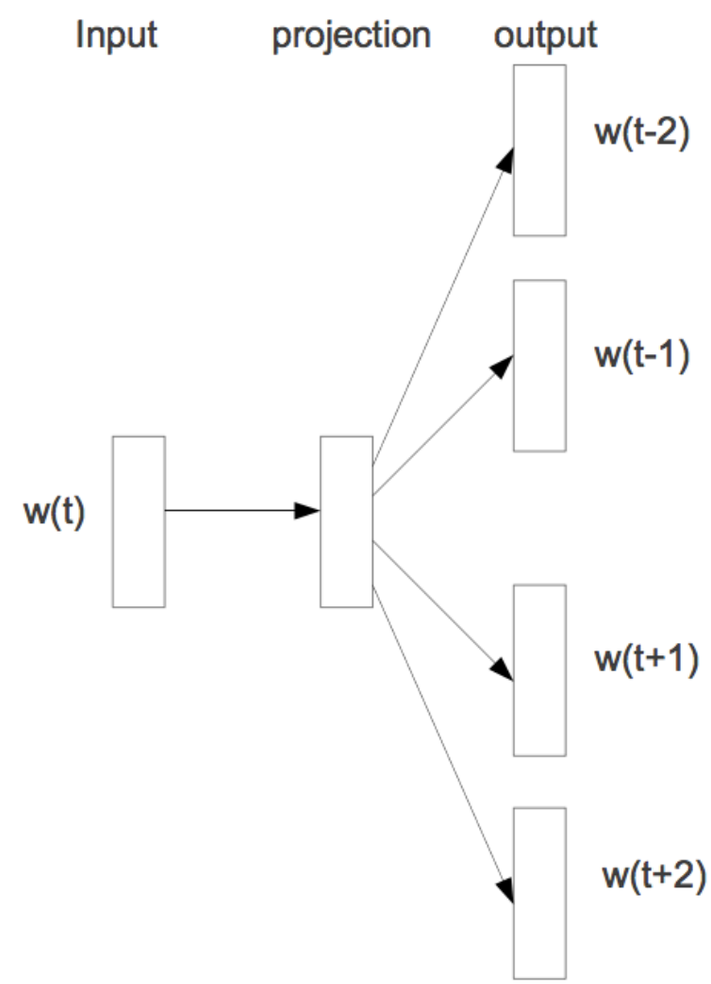
- ニューラルネットワークを用いた単語分散表現手法
- Skip-gram（中心単語から周囲に出現する単語を予測）を適用

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

最大になるよう重みを学習

$$p(w_{t+j}|w_t) = \frac{\exp(v'^{\mathrm{T}}_{w_{t+j}} v_{w_t})}{\sum_{w=1}^{W}\exp(v'^{\mathrm{T}}_{w} v_{w_t})}$$

$w_t$：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）入力単語”where”，N=3の時

where→<wh, whe, her, ere, re>

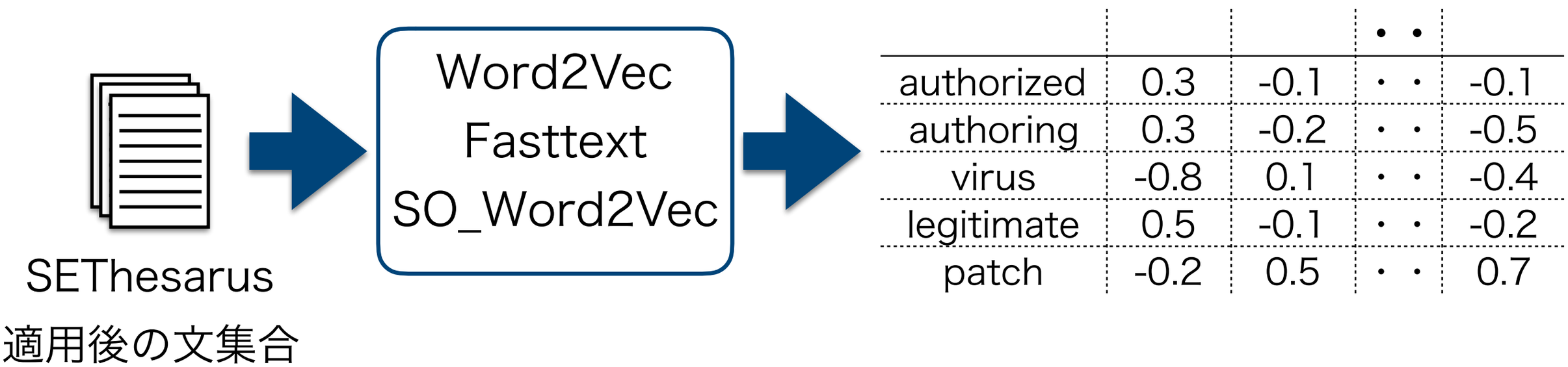
$$p(w_{t+j}|w_t) = \frac{\exp(v'^{\mathrm{T}}_{w_{t+j}} v_{w_t})}{\sum_{w=1}^{W} \exp(v'^{\mathrm{T}}_{w} v_{w_t})} \Rightarrow p(w_{t+j}|w_t) = \frac{exp(\sum_{g=1}^{G} z^{T}_{v_{w_t,g}} v_{w_{t+j}})}{\sum_{w=1}^{W} exp(v'^{T}_{w} v_{w_t})}$$

$z^{T}_{v_{w_t,g}}$：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

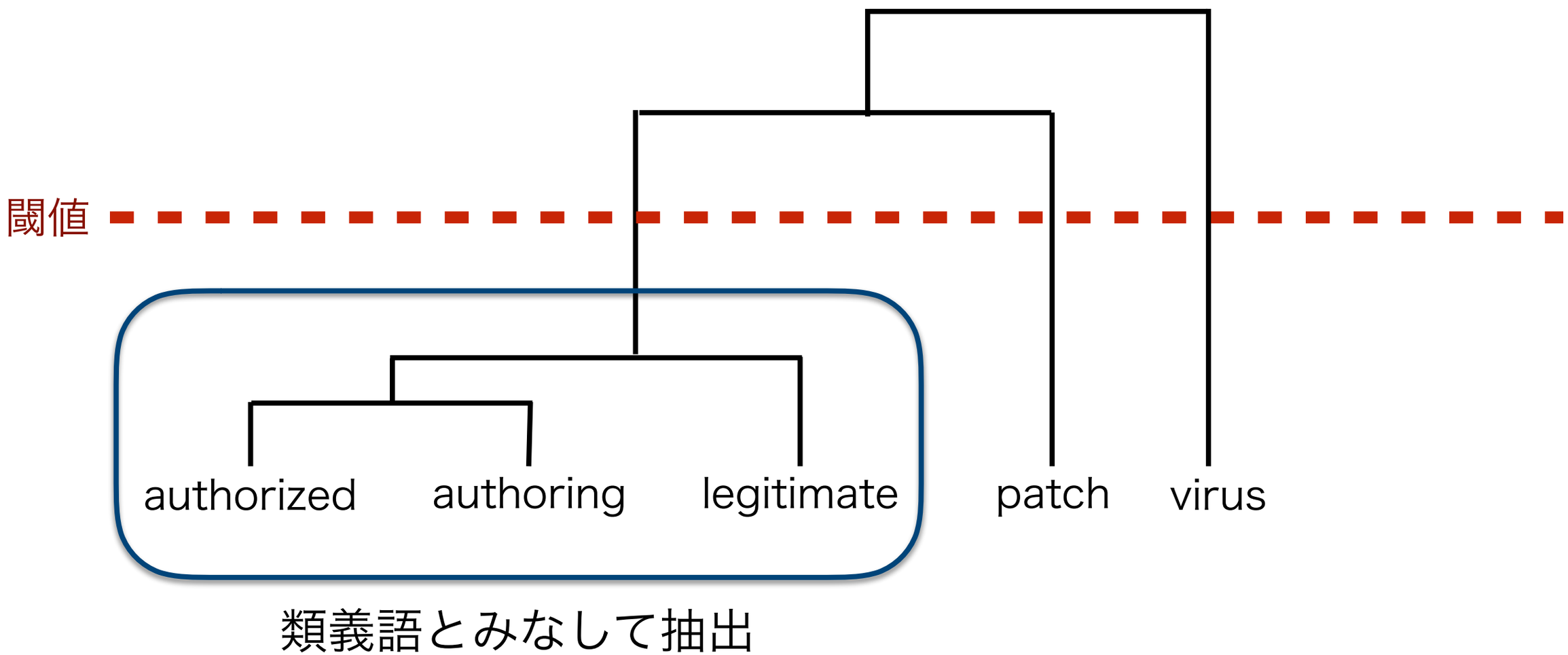
# 分散表現手法を用いた類義語の統一

- 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はSEThesaurus適用後の文集合を学習
    - 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



# 分散表現手法を用いた類義語の統一

- コサイン距離を用いて閾値以下で構成される単語クラスタを抽出
  - クラスタ距離の計算には外れ値に強い群平均法を適用



# 分散表現手法を用いた類義語の統一

- 単語クラスタの構成単語の一つに単語を統一

クラスタA

authorized
authoring
legitimate

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

統一

- overridden by **クラスタA** users
- **クラスタA** HTML documents
- impersonating a **クラスタA** user

# 提案手法の概要

前処理

↓

シソーラスを用いた
類義語の統一

↓

分散表現を用いた
類義語の統一

↓

**N-gramを特徴量とした
特徴ベクトルの作成**

↓

分類器の構築

# N-gramを特徴量とした特徴ベクトルの作成

1. N-gram（N=2,3）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. N-gram（N≧1）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 |

# 全てのN-gram（N≧1）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

$$MED(g) = log\frac{D(w_1, \cdots, w_n)}{D(g)}$$

$D(w_1, \cdots, w_n)$: Nグラムgを構成する単語を全て含む文書数

$D(g)$: Nグラムgを含む文書数

$t_1 t_2 t_3$

$t_3 t_1 t_2$

$t_1 t_2$

$$MED(t_1 t_2) = log\frac{3}{3} = 0$$

$$MED(t_1 t_2 t_3) = log\frac{2}{1} = log2$$

[10]Fan Bu, Xiaoyan Zhu, and Ming Li. 2010. Measuring the non-compositionality of multiword expressions. In Proceedings of the 23rd International Conference on Computational Linguistics (COLING '10). Association for Computational Linguistics, Stroudsburg, PA, USA, 116-124.

# 提案手法の概要

前処理

↓

シソーラスを用いた
類義語の統一

↓

分散表現を用いた
類義語の統一

↓

N-gramを特徴量とした
特徴ベクトルの作成

↓

**分類器の構築**

# 分類器の構築

- 提案手法と同じナイーブベイズとSMO，関連研究で用いられているSVMを適用
  - SMO：SVMの最適化問題を解くためのアルゴリズム

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

ナイーブベイズ
SMO
SVM

# 評価実験

- 既存手法と以下の組み合わせを比較

| | SEThesaurus | 分散表現手法 |
|---|---|---|
| 1 | あり | Word2Vec |
| 2 | あり | Fasttext |
| 3 | あり | SO_Word2vec |
| 4 | なし | SO_Word2vec |

×

| 特徴ベクトル |
|---|
| 1-gram |
| 2-gram |
| 3-gram |
| N-gram |

Word2Vec・Fasttextのパラメータ

| 次元数 | 最小出現回数 | ウィンドウ幅 | ネガティブサンプリング |
|---|---|---|---|
| 200 | 15 | 10 | 5 |

- 評価指標：Precision・Recall・F値
  - 対応のあるt検定を行い統計的有意差も確認

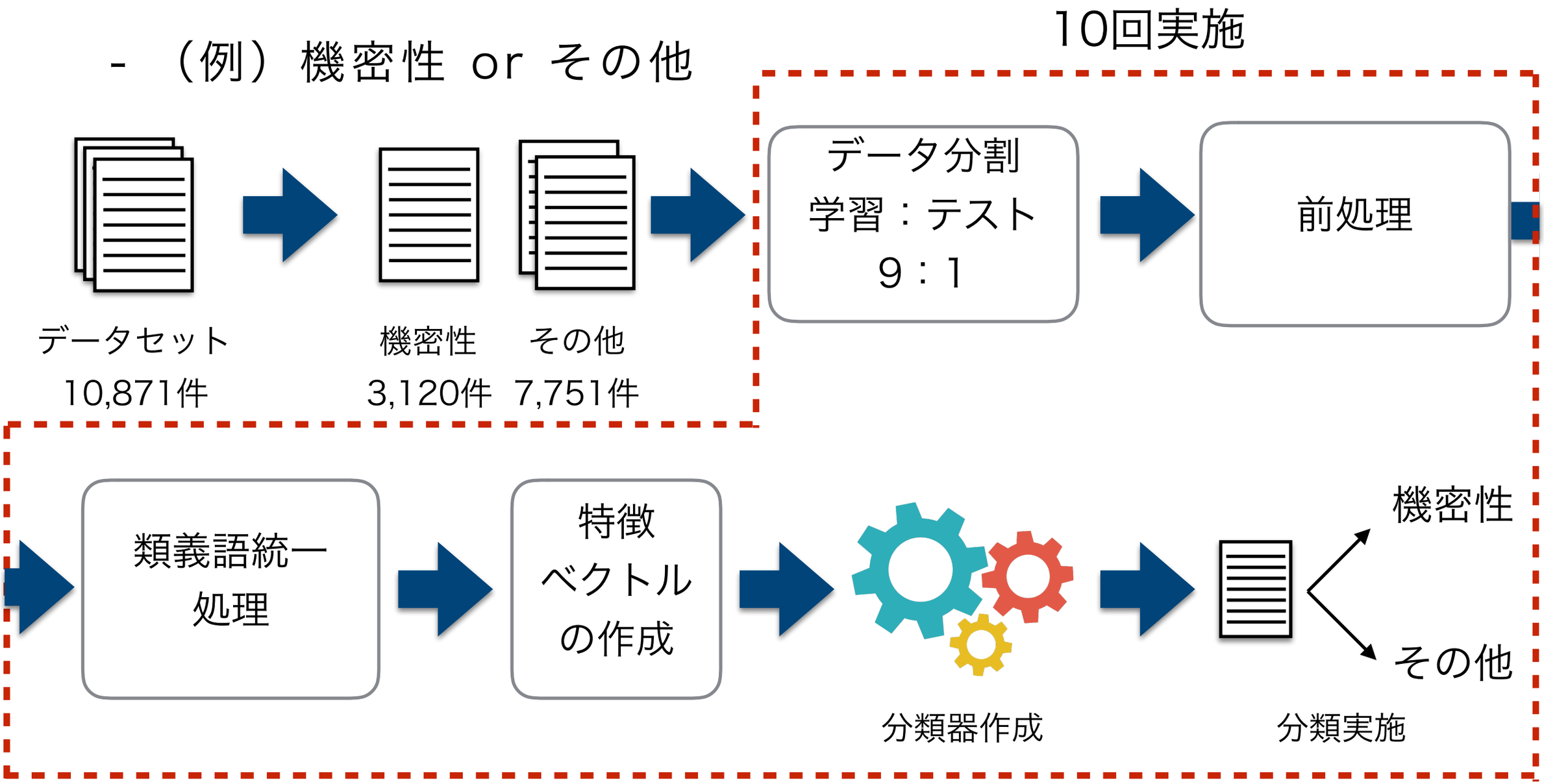
# データセット

- 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 | 概要 | 件数 |
|---|---|---|
| **機密性** | **許可された主体だけがデータにアクセスできるための度合い** | **3,120** |
| **完全性** | **不正な悪意のあるデータの修正や破壊を防止するための度合い** | **3,386** |
| 可用性 | ユーザーがいつでもシステムを利用できるための度合い | 174 |
| **責任追跡性** | **ユーザーやデータなどの行動を後から追跡できる度合い** | **3,882** |
| 真正性 | ユーザーやデータの確実に認識するための度合い | 311 |
| プライバシー | ユーザーがデータの使用方法を理解し制御するための度合い | 245 |
| それ以外 | 機能要件およびセキュリティ要件以外の要件文 | 6,617 |
| | total | 10,871 |

# 実験方法

- 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10回実施
  - （例）機密性 or その他



# 実験結果

- 機密性の分類結果

| 分類器 | 提案手法 | | | | | | 既存手法 | | | p値 |
|---|---|---|---|---|---|---|---|---|---|---|
| | SEThesaurus | 分散表現（閾値） | 特徴ベクトル | Pre | Re | F値 | Pre | Re | F値 | |
| ナイーブベイズ | あり | SO_W2V (0.14) | 2-gram | 0.751 | **0.807** | 0.778 | 0.709 | 0.783 | 0.744 | 0.000 |
| SVM | あり | SO_W2V (0.02) | N-gram | 0.809 | 0.770 | **0.789** | 0.778 | 0.748 | 0.762 | 0.000 |
| SMO | なし | SO_W2V (0.02) | N-gram | **0.810** | 0.766 | 0.787 | 0.789 | 0.739 | 0.763 | 0.000 |

Pre:Precision

Re:Recall

# 実験結果

・完全性の分類結果

| 分類器 | 提案手法 | | | | | | 既存手法 | | | p値 |
|---|---|---|---|---|---|---|---|---|---|---|
| | SEThesaurus | 分散表現（閾値） | 特徴ベクトル | Pre | Re | F値 | Pre | Re | F値 | |
| ナイーブベイズ | あり | Fasttext (0.02) | 2-gram | 0.742 | **0.829** | 0.783 | 0.708 | 0.806 | 0.754 | 0.000 |
| SVM | なし | SO_W2V (0.12) | N-gram | **0.806** | 0.781 | **0.793** | 0.782 | 0.767 | 0.774 | 0.000 |
| SMO | なし | SO_W2V (0.14) | N-gram | 0.805 | 0.776 | 0.790 | 0.790 | 0.762 | 0.775 | 0.000 |

Pre:Precision

Re:Recall

# 実験結果

・責任追跡性の分類結果

| 分類器 | 提案手法 | | | | | | 既存手法 | | | p値 |
|---|---|---|---|---|---|---|---|---|---|---|
| | SEThesaurus | 分散表現(閾値) | 特徴ベクトル | Pre | Re | F値 | Pre | Re | F値 | |
| ナイーブベイズ | あり | SO_W2V (0.14) | 2-gram | 0.835 | 0.834 | 0.834 | 0.769 | **0.855** | 0.810 | 0.000 |
| SVM | なし | SO_W2V (0.18) | N-gram | **0.856** | 0.832 | **0.844** | 0.829 | 0.817 | 0.823 | 0.000 |
| SMO | なし | SO_W2V (0.18) | N-gram | 0.853 | 0.826 | 0.839 | 0.840 | 0.810 | 0.825 | 0.000 |

Pre:Precision

Re:Recall

# 考察：MEDを用いたN-gramの抽出

| | フレーズ | 概要 |
|---|---|---|
| 非機能要件に関連するフレーズ | system shall provide | ～するものとする |
| | must capable of | ～しなければならない |
| セキュリティに関連するフレーズ | audit log | 監査のためのログ |
| | patient id | 認証のためのID |
| | log in | 認証に関係 |

- MEDを適用することで，ドメインで出現するフレーズを抽出することが可能

# まとめ

- セキュリティ要件の分類精度向上のために以下の処理を追加
  - SEThesaurusと分散表現を用いた類義語の統一処理
  - N-gramを特徴量とした特徴ベクトルの構築
- 機密性，完全性では精度の向上を確認