

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重要視する品質特性は異なる
 - 自動運転システムでは機能適合性が、金融系のシステムではセキュリティが重要視される

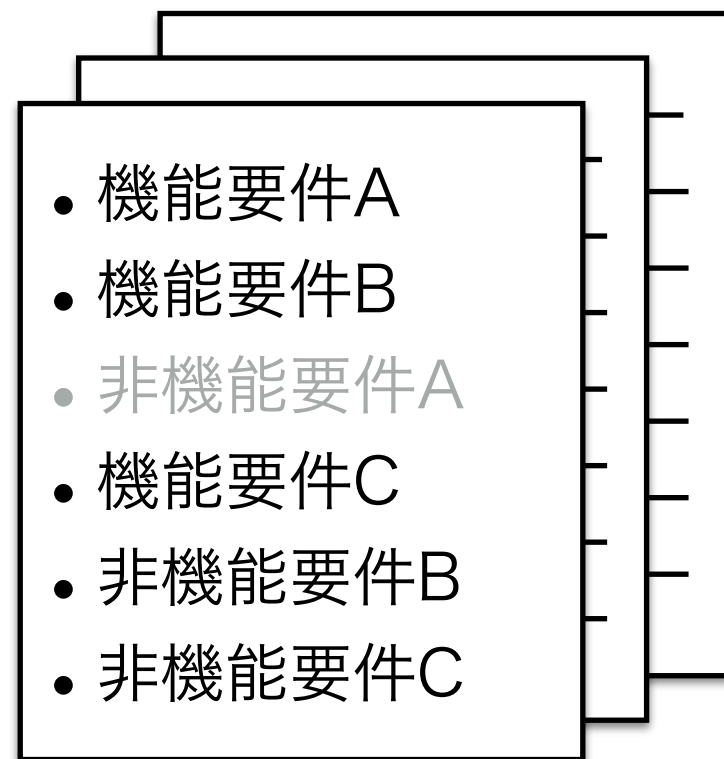
ISO25010[1]における品質特性

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	

- 開発するシステムが満たすべき品質特性を把握する必要がある

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



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]
 - ▶ 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]L.M. Cysneiros, Evaluating the effectiveness of using catalogues to elicit non functional requirements,” Workshop em Engenharia de Requisitos (WER '07, pp.107–115, 2007.

[3]A. Borg, A. Yong, P. Carlshamre, and K. Sandahl, “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.

非機要件分類手法[4,5]

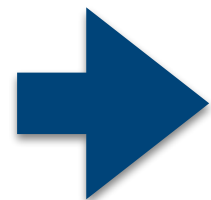
- ナイーブベイズ, SMO, SVMを用いた非機要件分類手法が提案されている

学習データ

(1)	Allows role-based access to EHR	機密性
(2)	Provides timely access to informa	機密性
(3)	Data used during user authentica	真正性
(4)	D2D performs identity managemen	真正性

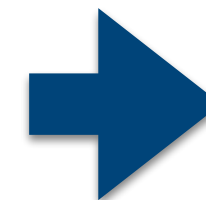


- 記号・ストップワードの除去
- 単語の正規化



	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

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]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.

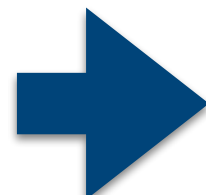
既存手法の問題点

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フレーズに対応できない

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

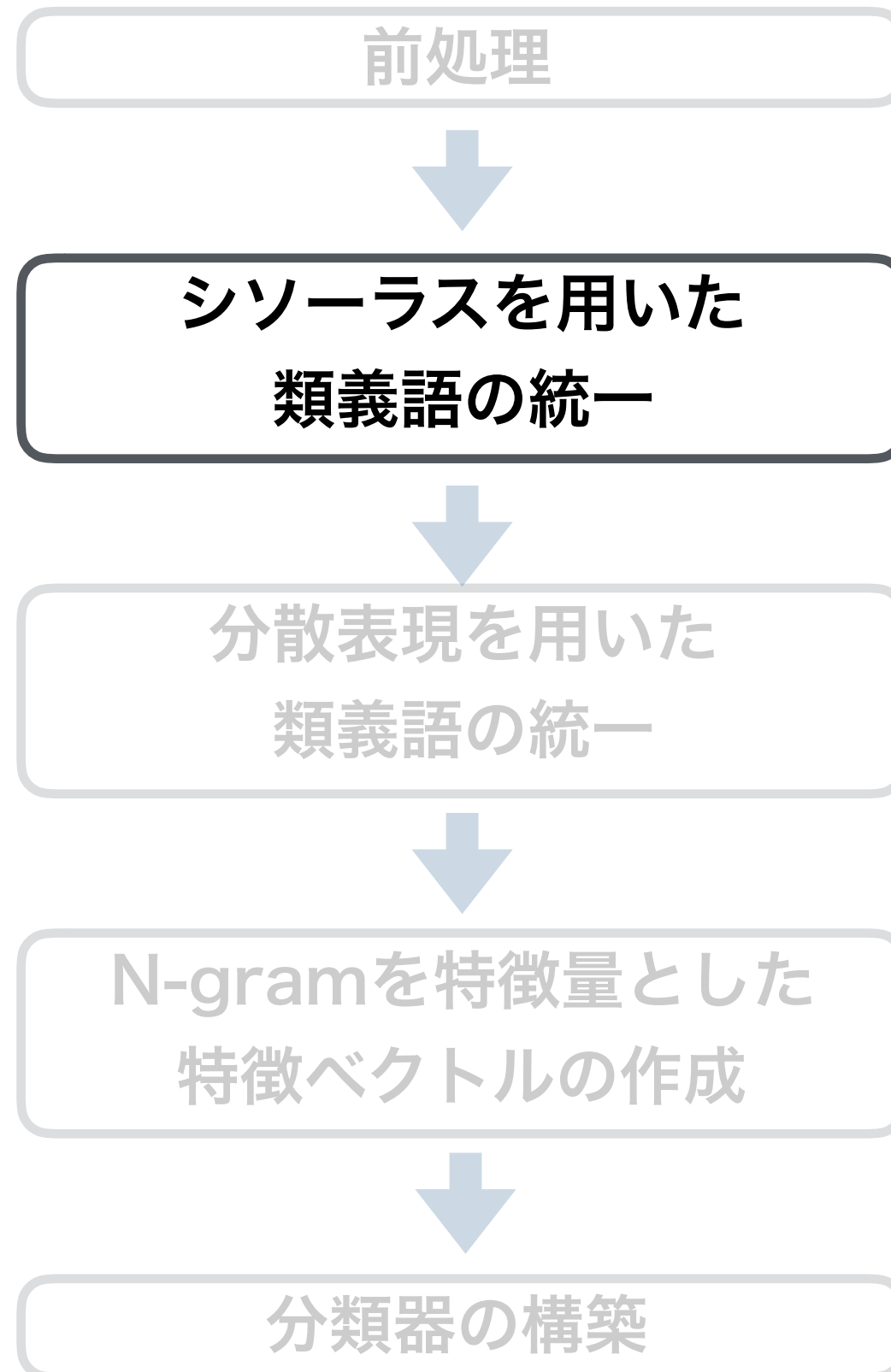
- ピリオドやカンマといった記号の除去
- 文字を大文字から小文字に変換
- 単語に分割
 - 数字を一つの表記に変換
 - ストップワードの除去
 - レマタイズ処理
 - 単語を基本形に変換する処理：（例）複数形を単数形に変換

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

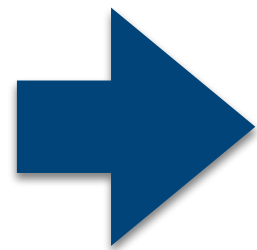


シソーラスを用いた類義語の統一

- 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Damerau-Levenshtein距離を用いて類義語を抽出
 - 分散表現の類似度が高く，文字列の類似度が高い単語

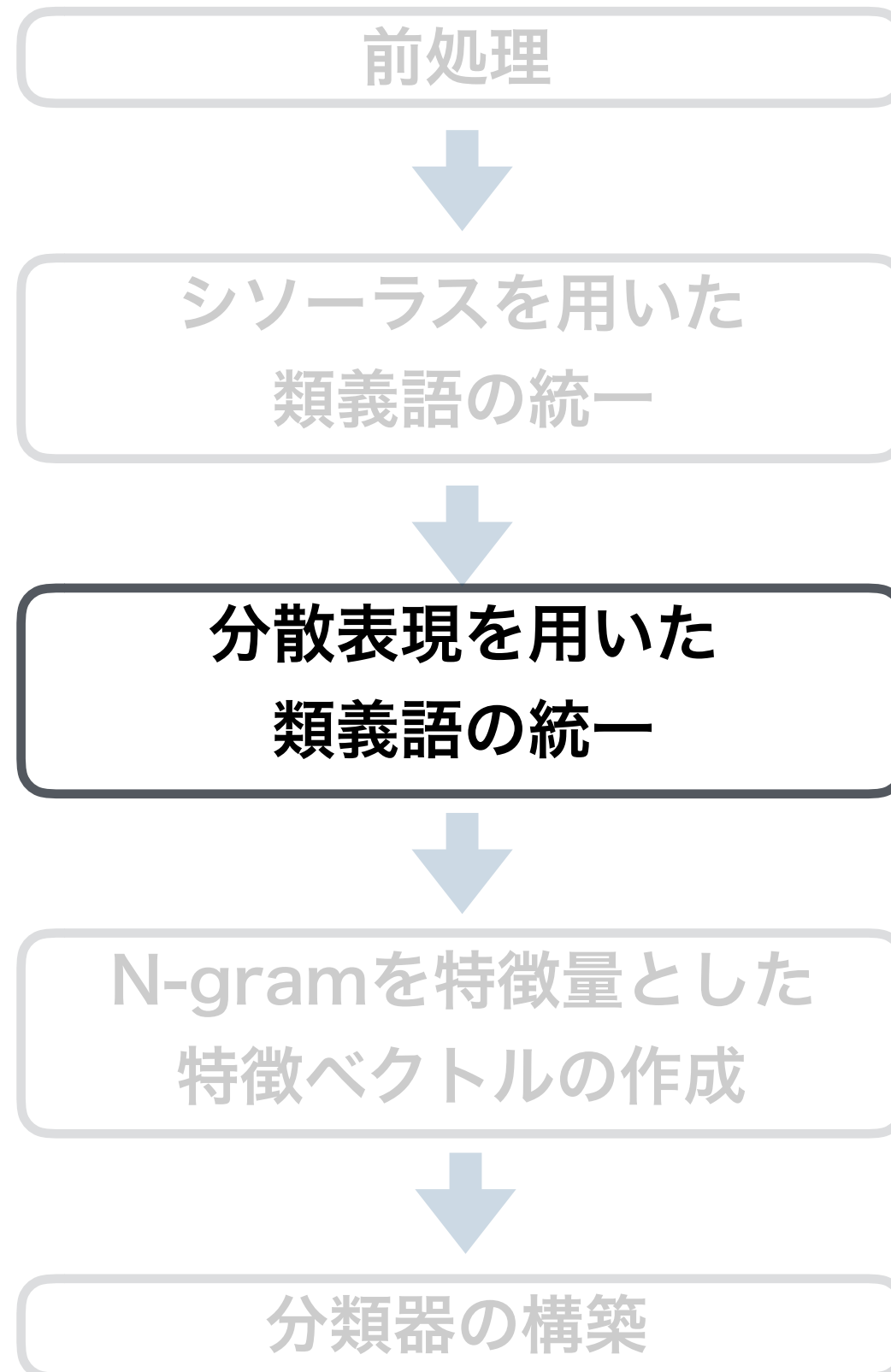
The system need be developed with **js.** ← SETheSaurus

- javascript
- javasript
- js



The system is developed in **Javascript.**

提案手法の概要



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 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

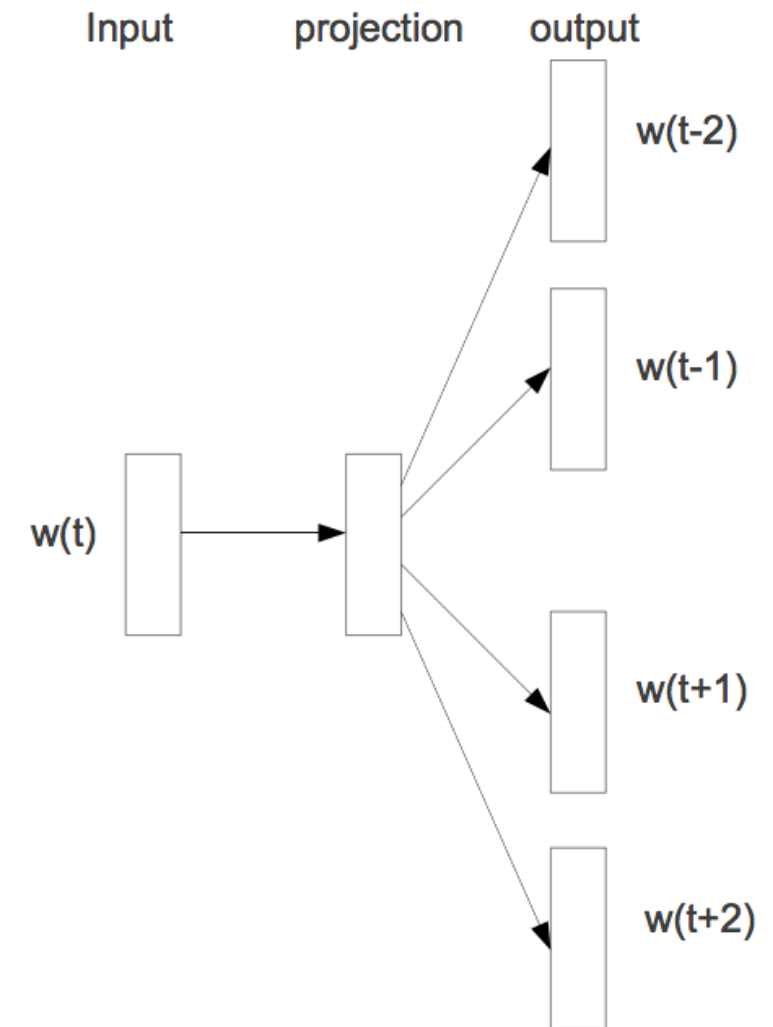
$$p(w_{t+j} | w_t) = \frac{\exp(v_{w_{t+j}}'^T v_{w_t})}{\sum_{w=1}^W \exp(v_w'^T v_{w_t})}$$

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-gramに分割
 - 文字Ngramのベクトルと出現単語のベクトルの内積の和を用いて出現確率を計算

例) 入力単語"where", N=3の時

where→<wh, whe, her, ere, re>

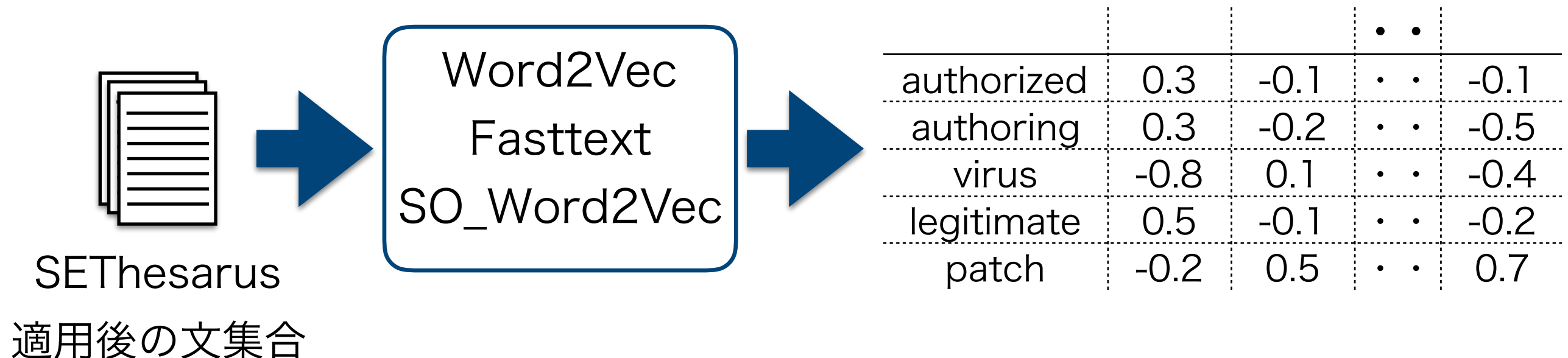
$$p(w_{t+j}|w_t) = \frac{\exp(v_{w_{t+j}}'^T v_{w_t})}{\sum_{w=1}^W \exp(v_w'^T v_{w_t})} \quad \rightarrow \quad p(w_{t+j}|w_t) = \frac{\exp(\sum_{g=1}^G z_{v_{w_t},g}^T v_{w_{t+j}})}{\sum_{w=1}^W \exp(v_w'^T v_{w_t})}$$

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

G : 単語をN-gramに分割してできた文字列の総数

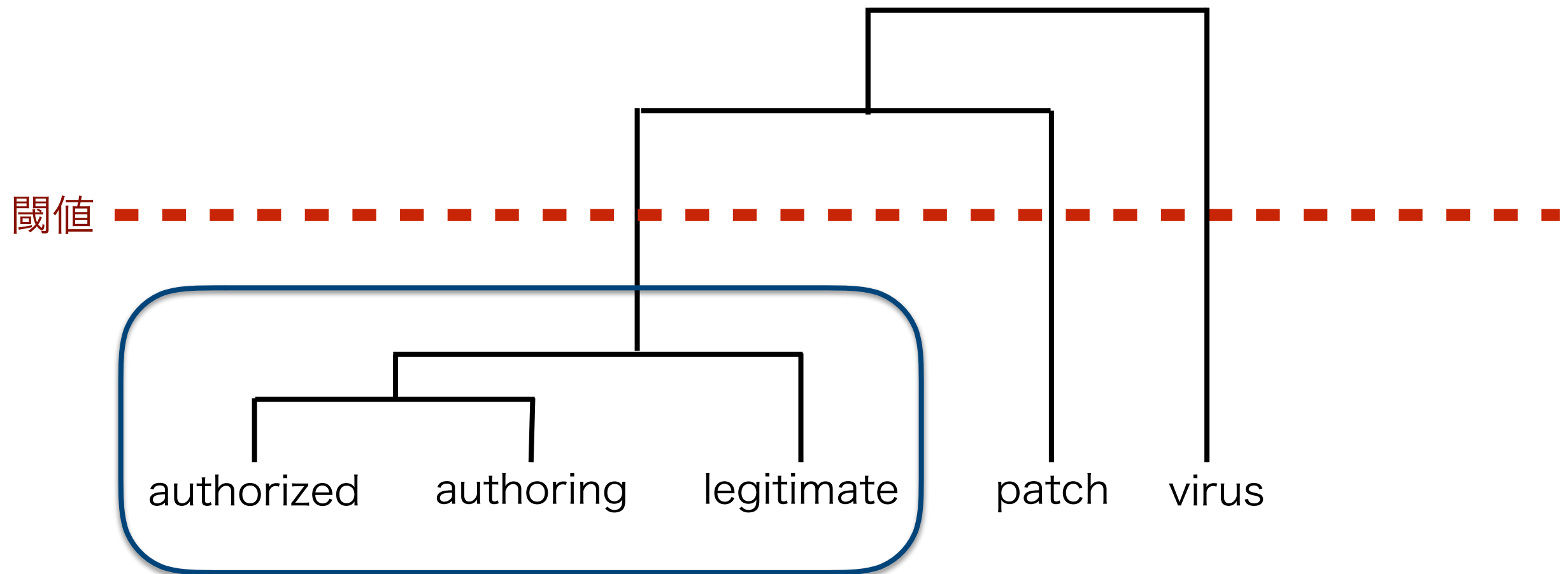
分散表現手法を用いた類義語の統一

- 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



類義語とみなして抽出

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

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

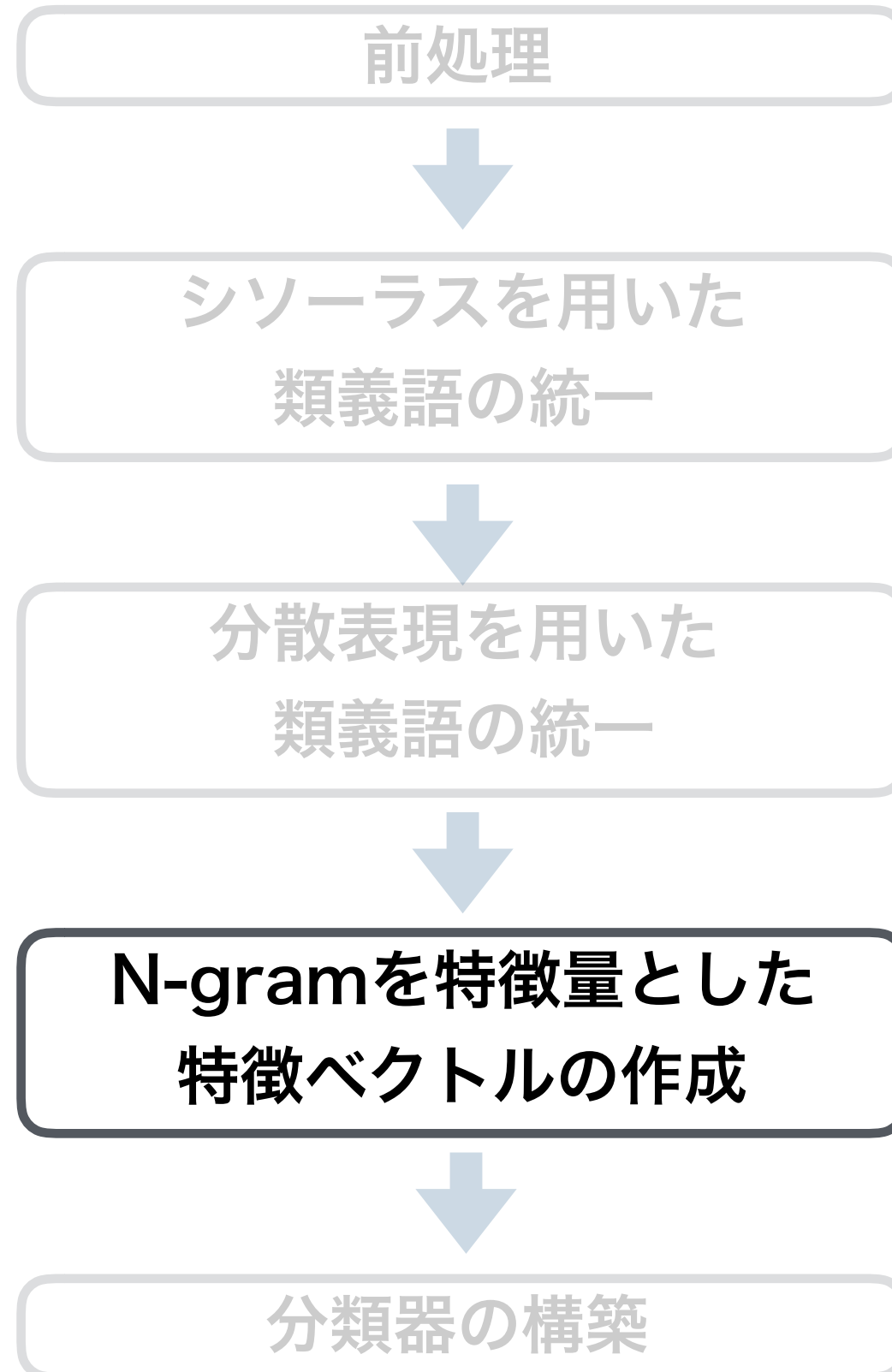
クラスタA

authorized
authoring
legitimate

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



N-gramを特徴量とした特徴ベクトルの作成

1. Bag-of-N-gram (N=2,3) 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 $\geq$ 1) を用いたN-gram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

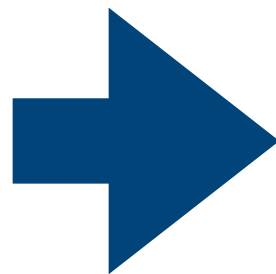
全てのN-gram (N≧1) を特徴量とした特徴ベクトルの作成

- 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

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

$D(w_1, \dots, w_n)$: 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

$t_1 t_2 t_3$
 $t_3 t_1 t_2$
 $t_1 t_2$



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

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

提案手法の概要



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	in as	log in as	・ ・	audit log
(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

- 既存手法と以下の類義語処理と特徴ベクトル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		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を確認

データセット

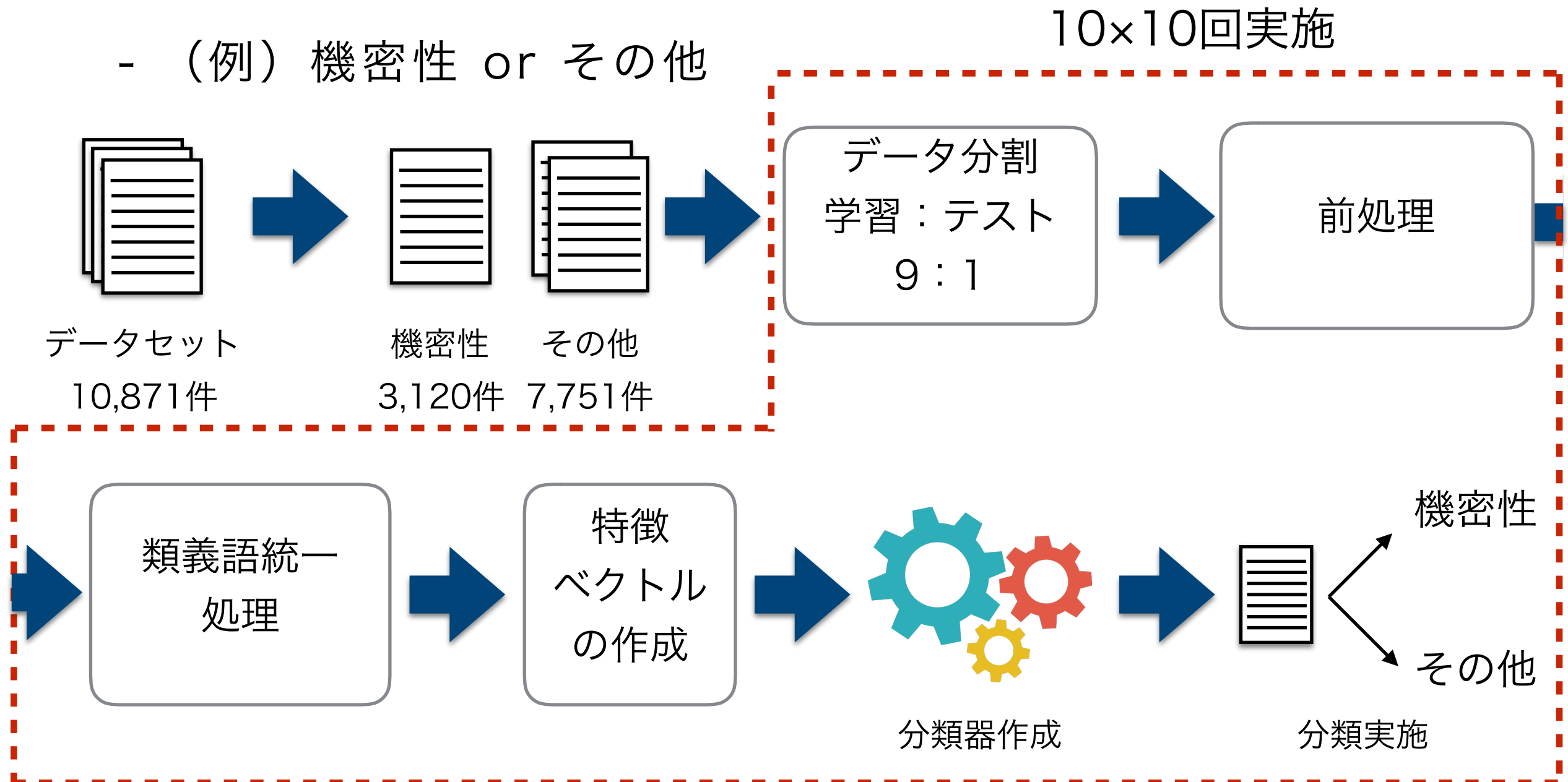
- 6つの医療系システムの要件定義書・機能要求
 - 1つの文に複数のセキュリティ要件が含まれる場合あり
 - データ数が一定以上ある機密性，完全性，責任追跡性を対象

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
合計		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（合計100回）

- （例）機密性 or その他



実験結果

- 機密性を対象にした分類結果

分類器	提案手法						既存手法		
	SE Thesaurus	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	0.7509	0.8067	0.7775	0.7085	0.7829	0.7436
SVM	あり	SO_W2V (0.02)	N-gram	0.8090	0.7696	0.7885	0.7779	0.7483	0.7625
SMO	なし	SO_W2V (0.02)	N-gram	0.8100	0.7661	0.7872	0.7889	0.7394	0.7630

- F値が約2.4～3.4%向上
 - p値<0.05を確認

実験結果

- 完全性を対象にした分類結果

分類器	提案手法						既存手法		
	SE Thesaurus	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	0.7419	0.8293	0.7830	0.7085	0.8058	0.7538
SVM	なし	SO_W2V (0.12)	N-gram	0.8059	0.7806	0.7928	0.7820	0.7668	0.7740
SMO	なし	SO_W2V (0.14)	N-gram	0.8052	0.7761	0.7901	0.7904	0.7615	0.7754

- F値が約1.5～2.9%向上
 - p値<0.05を確認

実験結果

- 責任追跡性を対象にした分類結果

分類器	提案手法						既存手法		
	SE Thesaurus	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	0.8348	0.8337	0.8341	0.7687	0.8553	0.8095
SVM	なし	SO_W2V (0.18)	N-gram	0.8559	0.8323	0.8438	0.8294	0.8166	0.8228
SMO	なし	SO_W2V (0.18)	N-gram	0.8535	0.8263	0.8395	0.8403	0.8099	0.8246

- F値が約1.5～2.5%向上
 - p値<0.05を確認
 - ただし、ナイーブベイズを用いた場合、Precisionは約6.6%向上したが、Recallは約2.2%低下

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	認証を指すフレーズ

- MEDを適用することでドメインのフレーズを抽出することが可能

まとめ

- セキュリティ要件分類精度向上のために以下の処理を追加
 - SEThesaurusと分散表現手法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
- 精度が良かった類義語処理と特徴ベクトルの組み合わせ結果
 - ナイーブベイズ2-gram特徴ベクトル, SVM・SMOではN-gram特徴ベクトル
 - ナイーブベイズの完全性以外でSO_Word2Vec
 - ただし, クラスタ距離はそれぞれ異なる