

Identifying Security Requirements Based on Linguistic Analysis and Machine Learning



Tong Li

2017 24th Asia-Pacific Software Engineering Conference
(APSEC), Nanjing, 2017, pp. 388-397.

背景

- セキュリティ侵害は年間数百万ドルもの損害を引き起こしている[1]
- セキュリティ要件を早期の段階で特定することで、開発コストを12~21%削減することが可能[2]
- 多くのプロジェクトではセキュリティの専門知識が不足しており、セキュリティ要件を特定するのは困難

[1]L. Ponemon, “Cost of data breach study: Global analysis,” Ponemon Institute sponsored by IBM, Tech. Rep., 2015.

[2]K. S. Hoo, A. W. Sudbury, and A. R. Jaquith, “Tangible roi through secure software engineering,” Secure Business Quarterly, vol. 1, no. 2, p. Q4, 2001.

関連研究

- 非機能要件を自動分類する手法が提案
 - セキュリティの特徴語を用いて分類する手法[3]
 - ナイーブベイズを用いて分類する手法[4]
- 統語論や意味論が考慮されていない
 - 過学習になってしまい，学習データのドメイン以外のシステムに適用した場合に精度が低下してしまう

[3] J. Cleland-Huang, R. Settimi, X. Zou, and P. Solc, “Automated classification of non-functional requirements,” *Requirements Engineering*, vol. 12, no. 2, pp. 103–120, 2007.

[4] E. Knauss, S. Houmb, K. Schneider, S. Islam, and J. Ju“rjens, “Supporting requirements engineers in recognising security issues,” in *International Working Conference on Requirements Engineering: Foundation for Software Quality*. Springer, 2011, pp. 4–18.

研究の目的・アプローチ

- 目的

- 文の意味構造を考慮したセキュリティ要件分類手法を構築すること

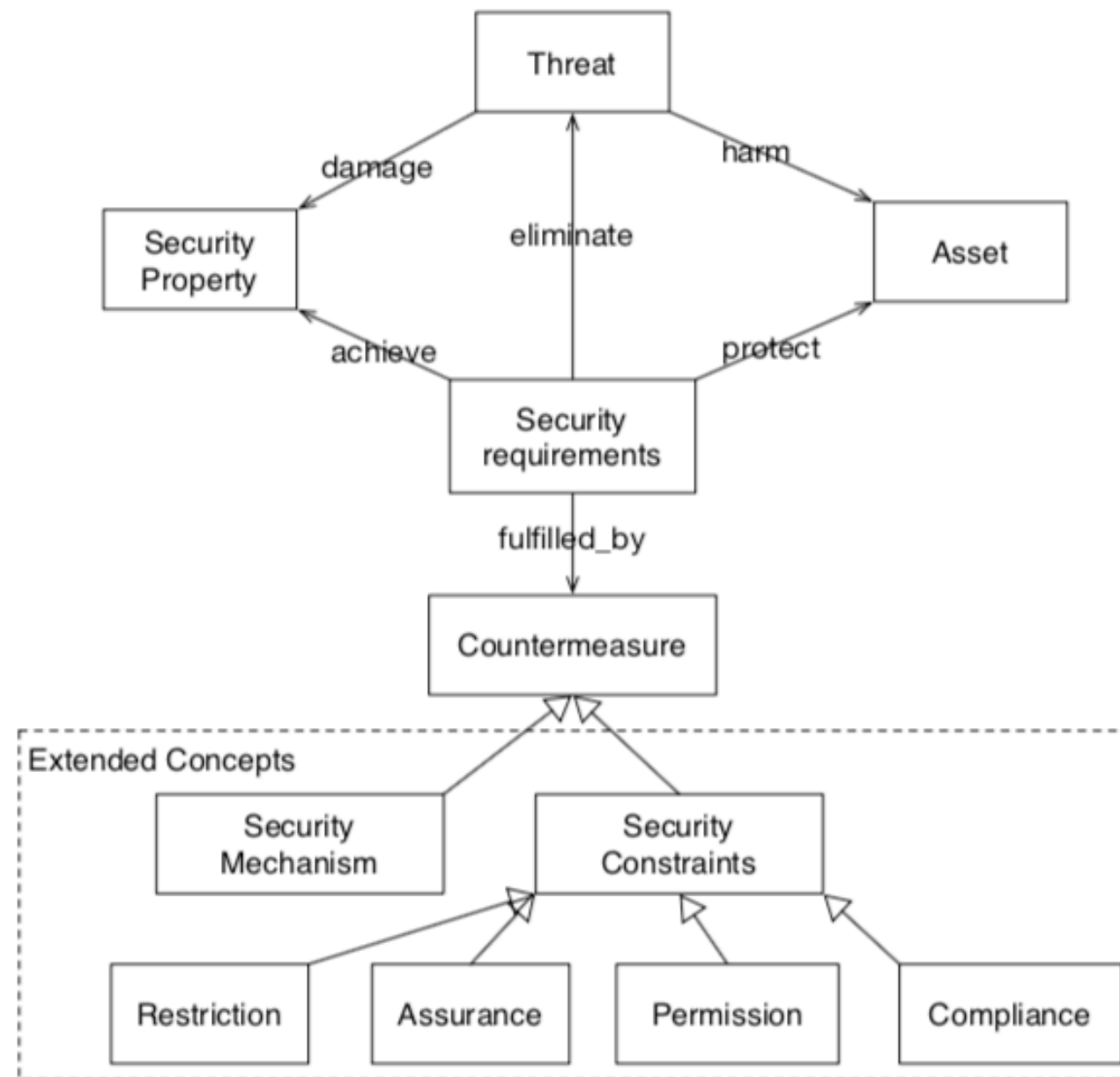
- アプローチ

- セキュリティ要件を示す記述ルールを定義
- ルールベースとキーワードベースを組み合わせた分類手法を構築

セキュリティ要件ルールの作成

.....

- 既存研究の概念モデル[5]を改変して定義



[5]D. Firesmith, "Specifying reusable security requirements," Journal of Object Technology, vol. 3, no. 1, pp. 61–75, 2004.

ルール1：Atomic Rule

.....

Name	Rules
Threat-based (T)	<i>< Subject >< Eliminate >< Threat ></i>
Asset-based (A)	<i>< Subject >< Protect >< Asset ></i>
Security property-based (S)	<i>< Subject >< Achieve >< SecurityProperty ></i>
	<i>< Subject >< Provide >< SecurityMechanism ></i>
	<i>< Subject >< Restrict > [< Object >]</i>
Countermeasure-based (C)	<i>< Subject >< Assure > [< SecureCondition >]</i>
	<i>< Subject >< Permit > [< SecureCondition >]</i>
	<i>< Subject >< Obey > [< SecurityPolicy >]</i>

- Threat-based：防ぐべき攻撃者が行う行為（情報窃盗など）を記した記述
- Asset-based：個人情報等の資産を保護することを示した記述
- Security property-based：機密性や完全性など達成すべきセキュリティ項目を示した記述
- Countermeasure-based：セキュリティを満たすために実施する対策を記した記述

ルール2：Compound Rule



Name	Rules
CA	<i>< Subject >< Provide >< SecurityMechanism > to < Protect >< Asset ></i>
CT	<i>< Subject >< Provide >< SecurityMechanism > to < Eliminate >< Threat ></i>
CS	<i>< Subject >< Provide >< SecurityMechanism > to < Achieve >< SecurityProperty ></i>
AT	<i>< Subject >< Protect >< Asset > to < Eliminate >< Threat ></i>
SA	<i>< Subject >< Protect >< SecurityProperty > of < Asset ></i>
TS	<i>< Subject >< Eliminate >< Threat > to < Achieve >< SecurityProperty ></i>
CAT	<i>< Subject >< Provide >< SecurityMechanism > to < Protect >< Asset > from < Threat ></i>
ATS	<i>< Subject >< Protect >< Asset > from < Threat > to < Achieve >< SecurityProperty ></i>
CSA	<i>< Subject >< Provide >< SecurityMechanism > to < Protect >< SecurityProperty > of < Asset ></i>
CTS	<i>< Subject >< Provide >< SecurityMechanism > to < Eliminate >< Threat > to < Achieve >< SecurityProperty ></i>
CATS	<i>< Subject >< Provide >< SecurityMechanism > to < Protect >< Asset > from < Threat > to < Achieve >< SecurityProperty ></i>

- Atomic Ruleを組み合わせたルール
 - 複数のセキュリティ要件が含まれている記述にも対応可能にするためのルール

セキュリティ要件キーワード

- 既存研究[6]やセキュリティ基準[7]に基づいて定義

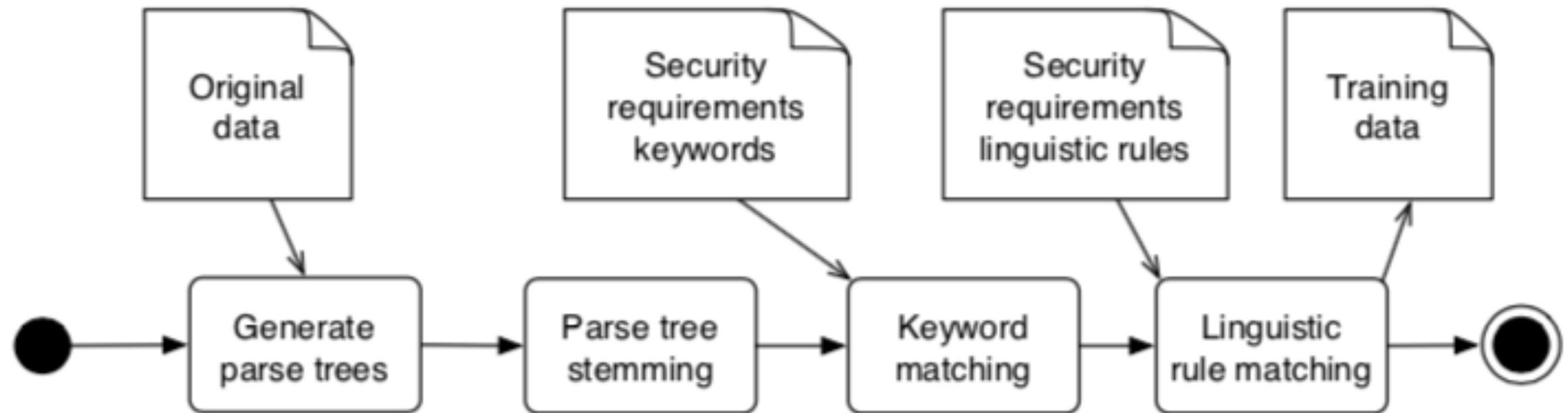
Entity	Keywords
Security Property	security, identification, authentication, authorization, integrity, immunity, anonymity, confidentiality, privacy, availability, non-repudiation, accountability, business continuity, reliability
Security Mechanism	auditing, access control, proof, key, lock, pin, detection, validation, verification, awareness, education, training, back-up, clear desk, clear screen, clock synchronization, cryptography, disposal, log, monitoring, segregation, separation, password, screening, protection, compliance, restriction, assurance, permission
Threat	fraud, error, bug, vulnerability, risk, fault, leakage, misuse, virus
Asset	NP
Subject	NP

Relation	Keywords
Eliminate	eliminate, get rid of, stop, reduce, avoid
Protect	protect, defend, save
Achieve	achieve, obtain, satisfy, fulfill
Provide	provide, implement, have, deploy
Restrict	restrict, limit
Assure	assure, ensure
Permit	permit, allow, admit
Obey	obey, comply, subject to

[6]D. G. Firesmith, “A taxonomy of security-related requirements,” in International Workshop on High Assurance Systems (RHAS’05). Citeseer, 2005.

[7] I. Committee, “Iso/iec 27000:2012 information technology – security techniques – information security management systems – overview and vocabulary,” Online: <http://www.27000.org/>, 2012.

提案手法の学習手順

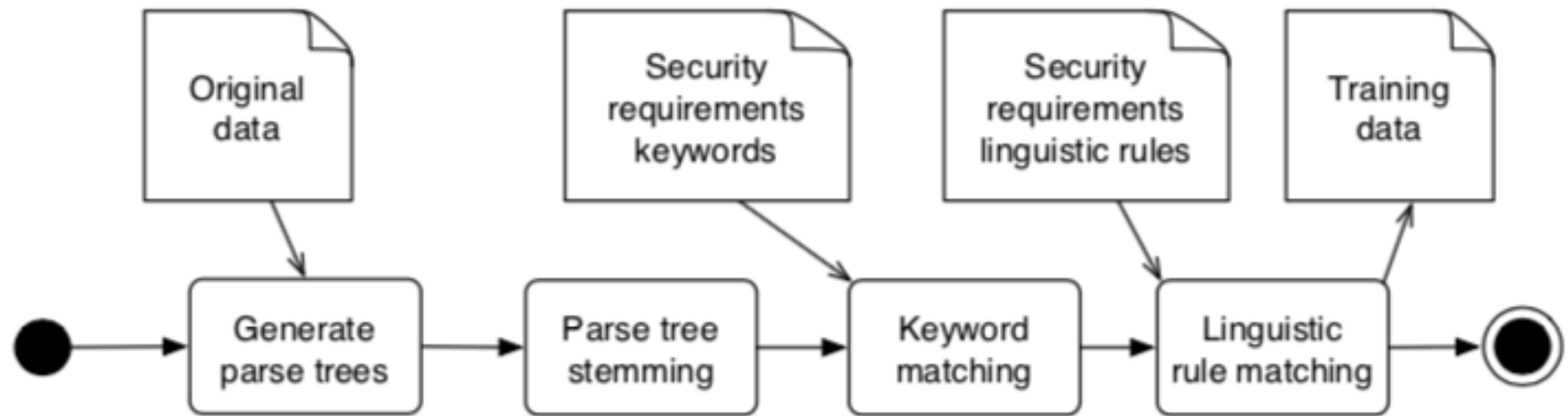


- 前処理

- 要求文を形態素解析し構文木を作成
- 単語をステミング処理をし，語幹に変換

提案手法の学習手順

.....



- 特徴ベクトルの作成

- Keyword matching

- セキュリティ要件キーワード含まれていれば1, そうでなければ0

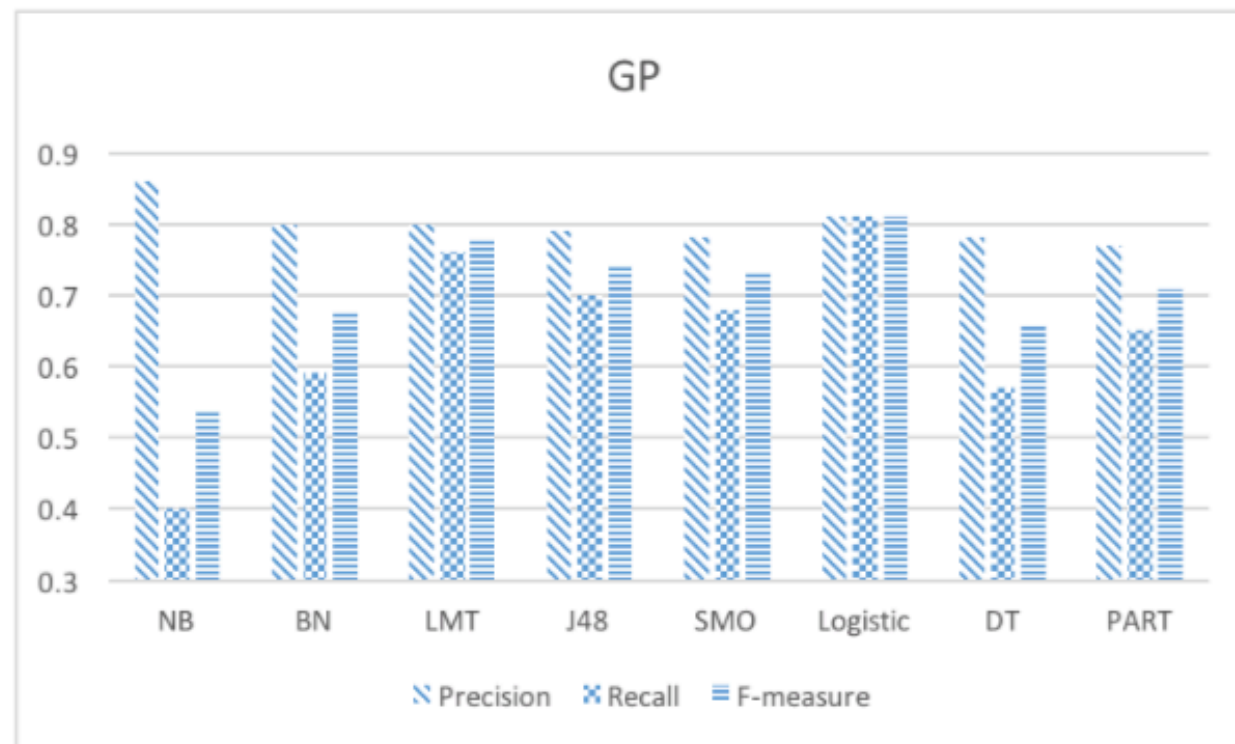
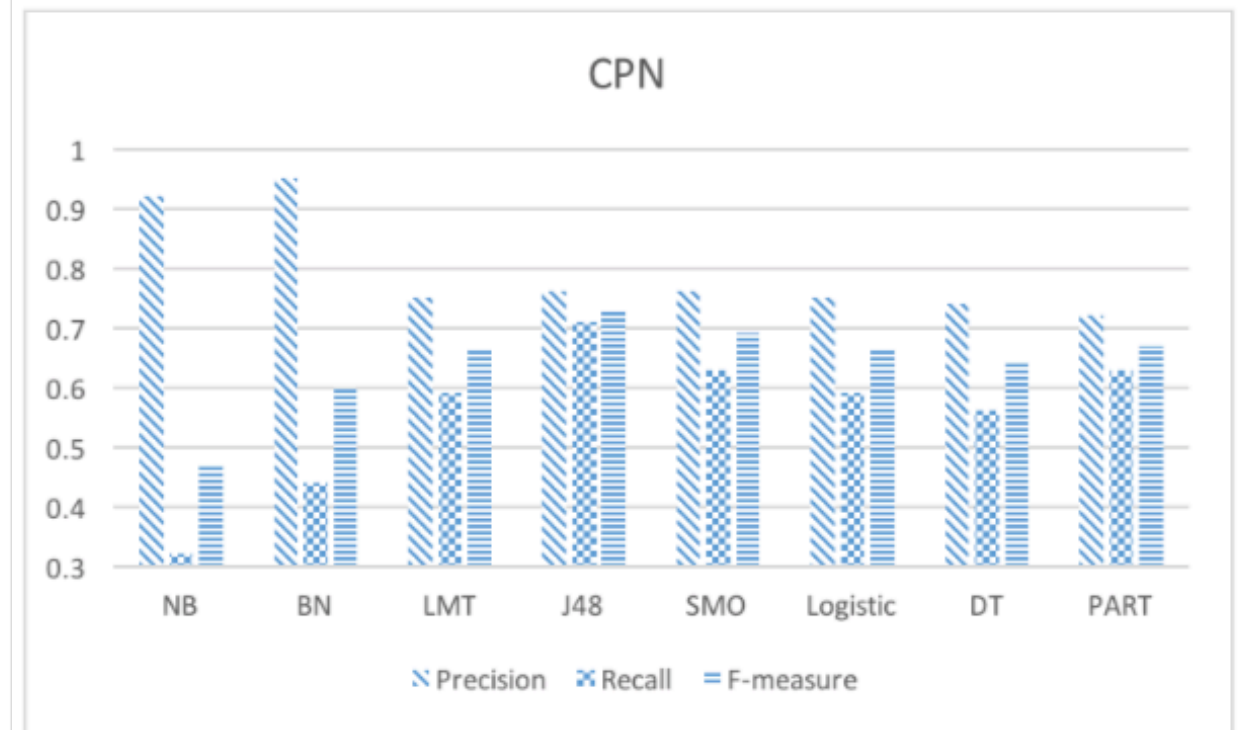
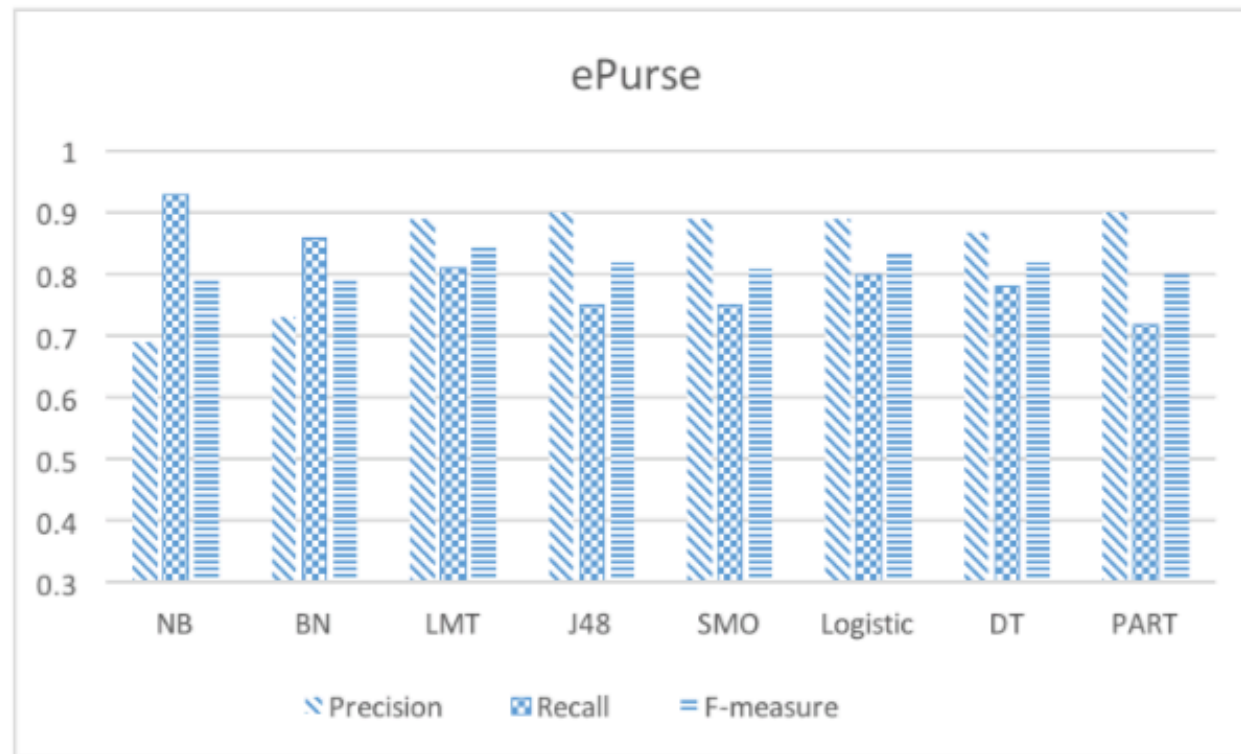
- Linguistic rule matching

- セキュリティ要件ルールの構文木と入力データの構文木を比較しマッチしていれば1, そうでなければ0とする

リサーチクエスチョン

- RQ1：最も精度の良い機械学習アルゴリズムはどれ？
 - ナイーブベイズ，ベイズネット，ロジスティックモデル木，J48，SVM，ロジスティック回帰，決定表，PARTを比較
 - ▶ ベイズの定理に基づいた手法，決定木の手法，関数を用いた手法，ルールベースの手法からそれぞれ2つずつ
- RQ2：提案手法はどの程度精度を向上するか？
- RQ3：提案手法は他のドメインのセキュリティ要件の分類に対応可能か？
 - 学習データとテストデータが違うデータセット

RQ1：最も精度の良い機械学習アルゴリズムはどれか？



- ・ ベイズ定理の手法が精度悪く他は同等の精度
- ・ Recallの最も高いJ48を採用

リサーチクエスチョン

- RQ1：最も精度の良い機械学習アルゴリズムはどれ？
 - ナイーブベイズ，ベイズネット，ロジスティックモデル木，J48，SVM，ロジスティック回帰，決定表，PARTを比較
 - ベイズの定理に基づいた手法，決定木の手法，関数を用いた手法，ルールベースの手法からそれぞれ2つづつ
- RQ2：提案手法はどの程度精度を向上するか？
- RQ3：提案手法は他のドメインのセキュリティ要件の分類に対応可能か？
 - 学習データとテストデータが違うデータセット

評価実験

.....

- データセット

Specification	Requirements	Security requirements	Percentage
ePurse	124	83	67%
CPN	210	41	20%
GP	176	63	36%

- 既存手法（ナイーブベイズ用いた手法）と比較
 - 特徴ベクトル作成に使う単語は記述文に含まれる単語全て
 - 10交差検証を実施
 - PrecisionとRecallとF1値を用いる

RQ2：提案手法はどの程度精度を向上するか？

.....

Dataset	Precision	Recall	F-measure
ePurse	0.83/0.9	0.93/0.75	0.88/0.82
CPN	0.98/0.76	0.95/0.71	0.96/0.73
GP	0.81/0.79	0.92/0.7	0.86/0.74
ePurse+CPN	0.81/0.77	0.93/0.74	0.87/0.76
ePurse+GP	0.8/0.82	0.96/0.83	0.87/0.82
CPN+GP	0.82/0.71	0.87/0.72	0.85/0.71
ePurse+CPN+GP	0.79/0.76	0.91/0.8	0.84/0.78
average	0.83/0.79	0.92/0.75	0.88/0.77

Remark: benchmark results are presented first

既存/提案

- 既存手法のほうが精度が良かった
 - CPNでは違いが顕著に

RQ3：提案手法は他のドメインのセキュリティ要件の分類に対応可能か？

.....

Dataset	Precision	Recall	F-measure
ePurse(CPN)	0.23/0.52	0.54/0.78	0.33/0.63
ePurse(GP)	0.43/0.5	0.85/0.78	0.57/0.61
CPN(ePurse)	0.99/0.95	0.33/0.49	0.47/0.65
CPN(GP)	0.29/0.85	0.19/0.17	0.23/0.29
GP(ePurse)	0.72/0.94	0.48/0.82	0.58/0.88
GP(CPN)	0.29/0.5	0.65/0.76	0.4/0.6
ePurse+CPN(GP)	0.51/0.52	0.56/0.43	0.53/0.47
ePurse+GP(CPN)	0.26/0.47	0.85/0.76	0.4/0.58
CPN+GP(ePurse)	0.84/0.95	0.31/0.75	0.46/0.84
average	0.51/0.69	0.53/0.64	0.44/0.61

Remark: benchmark results are presented first

既存/提案

- 提案手法の方が精度良い

考察

- RQ2では既存手法， RQ3では提案手法が良かった
 - 既存手法は学習データのドメインの単語パターンを学習しているため， 学習データと同じドメインのデータであれば分類可能
 - ドメインが異なると単語の出現頻度が変わってしまうので， 既存手法で学習データと異なるドメインの文进行分类するのは不向き

まとめ・所感

- まとめ

- 文の意味構造を考慮したセキュリティ要件分類手法を提案
- 提案手法の方がドメインに依存せずに分類することが可能

- 所感

- 意味構造ありなしの純粋な比較になっていない
 - 特徴ベクトル, 学習アルゴリズムが違う
 - 同じ特徴ベクトル・アルゴリズム場合の精度の影響が気になる