

非機能要件自動抽出手法の構築

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

- 近年のシステム開発ではベンダーに開発を委託する委託開発が主流
- ベンダーは要件定義を行い，開発するシステムに必要な要件を定義にする

要件	概要
機能要件	システムに実装する機能のこと
非機能要件	機能要件以外の要件 性能やセキュリティなど

非機能要件

.....

- ”非機能”の幅広い範囲を対象[1]

要件	概要
可用性	システムの継続性に関する項目
性能・拡張性	システムの性能や拡張性に関する項目
運用・保守性	運用・保守作業に関する項目
移行性	新システムへの移行に関する項目
セキュリティ	セキュリティに関する項目
システム環境・エコロジー	消費電力など自然環境に関する項目

[1]非機能要求グレード, <https://www.ipa.go.jp/sec/softwareengineering/reports/20100416.html>

セキュリティ要件の項目

大項目	小項目	メトリクス
セキュリティ	情報セキュリティに関する コンプライアンス	順守すべき社内規程、ルール、法令、 ガイドライン等の有無
	セキュリティリスク分析	リスク分析範囲
	セキュリティ診断	ネットワーク診断実施の有無
		Web診断実施の有無
		DB診断実施の有無
	セキュリティリスク の見直し	セキュリティリスク見直し頻度
		セキュリティリスクの見直し範囲
	セキュリティリスク対策の 見直し	運用開始後のリスク対応範囲
		リスク対策方針
	セキュリティパッチ適用	セキュリティパッチ適用範囲
		セキュリティパッチ適用方針
		セキュリティパッチ適用タイミング
	認証機能	管理権限を持つ主体の認証
		管理権限を持たない主体の認証
	利用制限	システム上の対策における操作制限度
		物理的な対策による操作制限度

セキュリティ要件の項目

大項目	小項目	メトリクス
セキュリティ	管理方法	管理ルールの策定
	データ暗号化	伝送データの暗号化の有無
		蓄積データの暗号化の有無
		鍵管理
	不正監視	ログの取得
		ログ保管期間
		不正監視対象（装置）
		不正監視対象（ネットワーク）
		不正監視対象（侵入者・不正操作等）
		確認間隔
	データ検証	デジタル署名の利用の有無
		確認間隔
	ネットワーク制御	通信制御
	不正検知	不正通信の検知範囲
	サービス停止攻撃の回避	ネットワークの輻輳対策

セキュリティ要件の項目

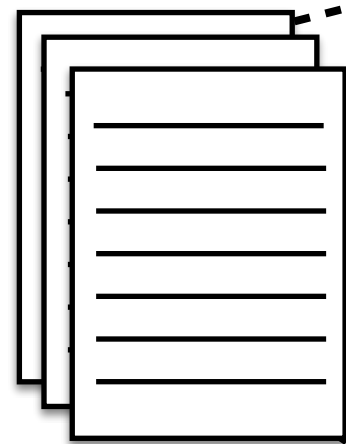
.....

大項目	小項目	メトリクス
セキュリティ	マルウェア対策	マルウェア対策実施範囲
		リアルタイムスキャンの実施
		フルスキャンの定期チェックタイミング
		セキュアコーディング、Webサーバの設定 等による対策の強化
	Web実装対策	WAFの導入の有無

- 他の非機能要件（可用性など）も同様に多くの項目で構成される
- ベンダーはこれらの中からシステムに必要な非機能要件を定義する

要件定義書

- システムの機能要件と非機能要件をまとめた文書
 - 多いもので約200ページほど
- 設計者やプログラマは要件定義書を元に設計・実装を行う
 - 別のベンダー会社に設計・実装を依頼する場合も



要件定義書

第1章 業務要件の定義

第1節 調達件名

厚生労働省情報提供システムの更改整備及び運用・保守業務一式

第2節 次期情報提供システムの対象業務の範囲

厚生労働省（以下、「当省」という。）では、現在運用中の厚生労働省ホームページを始め、厚生労働省法令等データベースサービス、厚生労働省図書館サービス等の様々なサービスを国民に提供している。現行情報提供システムは、平成25年7月にシステム更改を実施したが、次期情報提供システムでは、給水装置データベース、厚生労働省予算の支出状況検索サイト等のデータベースサイトや都道府県労働局情報提供サイト集約化システム等のポータルサイトを集約・統合し、「厚生労働省ホームページ等情報提供システム」、「都道府県労働局情報提供サイト集約化システム」、「地方厚生局Webサーバー」、「海外感染症情報システム」及び「共通基盤システム」として再構成してシステム更改を行い、情報提供システムの最適化を図る。

次期情報提供システムは、これまでの情報提供手段に加えて、オープンデータの拡充

機能要件の記述例

(1) 実装機能

厚生労働省ホームページ等情報提供システムの各業務処理機能の概要を「表 2-1 厚生労働省ホームページ等情報提供システムで提供する主な業務処理機能の概要」に示す。

表 2-1 厚生労働省ホームページ等情報提供システムで提供する主な業務処理機能の概要

項番	サービス	主な業務処理機能	概要
1	WWWサービス		
	厚生労働省ホームページ	Webサーバ	● 当省の組織、厚生労働行政における様々な情報（報道発表資料・緊急情報等）を配信する。
	ハンセン病療養所サイト		● 国立療養所（ハンセン病療養所 13 施設）の施設情報、診療案内等の様々な情報を広く国民へ情報を配信する。
	中央労働委員会労働委員会関係命令・裁判例データベース		● 中央労働委員会及び都道府県労働委員会が発した不当労働行為の命令書等及びその後に労働委員会等が当事者となった裁判所の判決等の情報を配信する。
	給水装置データベース		● 給水装置が水道法に基づく構造、材質基準の適合、認証情報等を水道利用者へ情報を配信する。
	厚生労働省予算の支出状況検索サイト		● 当省における予算の支出状況を管理し、情報を配信する。

項番	サービス	主な業務処理機能	概要
	厚生労働省法令等データベース		● 当省が所管する全ての法令・通知・公示情報を配信する。
	厚生労働省図書館ホームページ		● 国立国会図書館の当省支部図書館として位置づけられ、主に当省図書館所管の蔵書情報を配信する。
2	厚生労働省図書館管理サービス	利用者管理	● 厚生労働省図書館を利用する利用者の管理を行う。
		蔵書・貸出管理	● 厚生労働省図書館の蔵書管理や利用者への貸出管理を行う。
3	政府共通NW用図書・検索サービス	政府共通NW用図書・検索サーバ	● 国立国会図書館用分散型目録データベースへのデータ連携を行う。
4	マルチキャリア対応携帯Webサービス	Webサーバ	● 厚生労働省ホームページにおいて、緊急情報、組織連絡先等のコンテンツを携帯端末（フィーチャーフォン）より閲覧可能とするため、携帯端末（フィーチャーフォン）用テンプレートを用いて広く国民に情報提供を行う。
5	アクセシビリティ改善ソフトウェアサービス	音声読み上げ／文字拡大／画面配色切替え	● ホームページにおける情報提供のアクセシビリティ確保又は、バリアフリー（無障壁）なアクセスを可能とするため、視力の弱い人、パソコン初心者及び高齢者等を対象とし、ブラウザ環境を用いて情報提供を行う。
6	メール配信サービス	会員管理	● 配信登録された会員の管理を行う。
		配信機能	● 配信登録された会員に対して、ホームページサービス上で、更新された「新着情報」及び「緊急情報」をメール配信する。

- 機能要件は有無で判断ができる
 - 必要な機能はユーザー・ベンダーとも定義しやすい

非機能要件の記述例

(4) 不正監視

次期情報提供システムにおいて、不正行為に迅速に対処するため、以下に示す要件を遵守すること。

- 本調達に係る情報システムに対する不正アクセス、ウイルス・不正プログラム感染等、インターネットを経由する攻撃、不正等への対策機能を有すること。
- 通信回線を介して当省外（インターネット又は他府省）と送受信される通信内容を監視し、不正アクセスや不正侵入を検知及び通知する機能を備えること。
- サービス不能攻撃及び標的型攻撃の検知及び通知をする機能を備え、これらの攻撃に対し迅速な対応を行うこと。
- 現状想定している不正監視の内容は「表 3-21 不正監視」のとおりである。これらの不正監視を行い、検知及び通知された内容に対し迅速な対応を行うこと。なお、「表 3-21 不正監視」に記載がない項目であっても、「第10節 情報セキュリティに関する事項」に示す要件を満たすために必要な監視対象がある場合は、適宜対象を追加し監視すること。

セキュリティに関する内容

「迅速な」とはどれくらい？
具体的な日数を記述する場合も

- 機能要件と比べて特定が困難
 - 非機能要件が機能要件に混載[2]
 - 表現方法のバリエーションが多い

非機能要件特定の課題

- 大量の時間を必要とする問題
- 解釈を間違えてしまう問題
- 見落としてしまう問題
 - システム構築後に発覚する 경우가ほとんど
 - 開発工程の手戻りやシステム障害の原因

本研究は非機能要件の見落とし
防止に取り組む

教師あり学習を用いた既存手法[3,4]

- 非機能要件の見落としを未然に防ぐためにSVMやNaive Bayesを用いた手法が提案されている

データの暗号化をすること	セキュリティ
応答時間は2秒以内	性能
⋮	⋮



学習

大量の学習データが必要となり
学習データ作成にコストがかかる



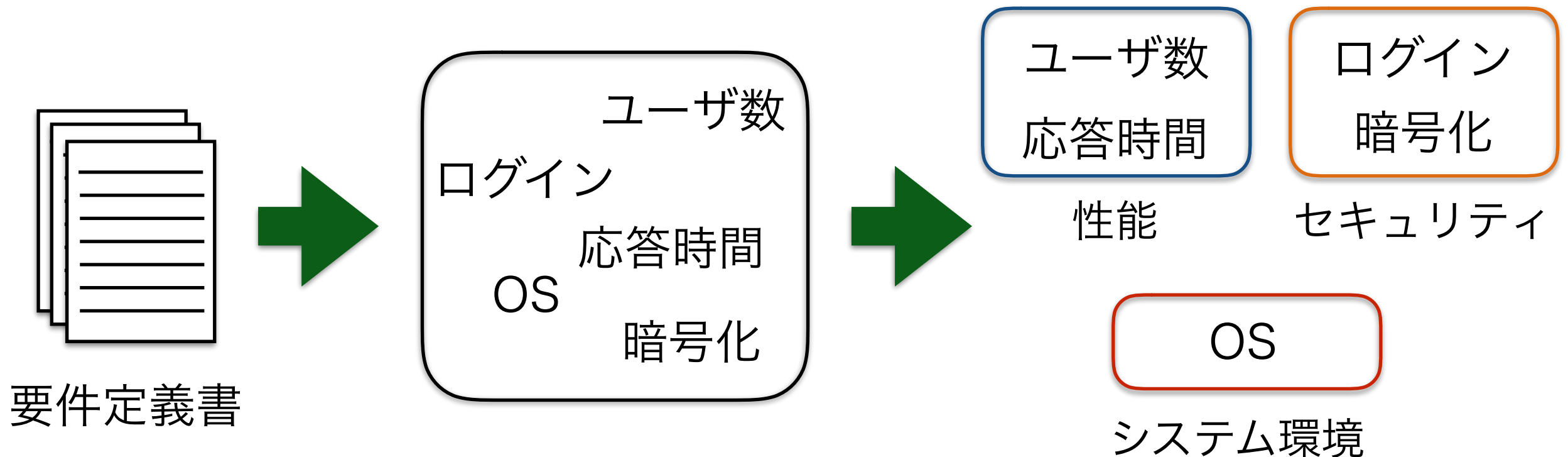
SVM等

[3]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)

[4]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[5]

- 要件定義書から各非機能要件に関連する単語を抽出し辞書を作成
 - Normalized Google Distanceと階層クラスタリングを適用
- 辞書と文の関連性を算出して分類を行う



Normalized Google Distance[6]

- 単語間の関連度を計算するアルゴリズム
 - コルモゴロフ複雑性をWeb検索で近似
 - 同じWebページ上で共起する頻度の多い単語は関連度が高いというアイデア

$$NGD(w_1, w_2) = \frac{\max\{\log(D_1), \log(D_2)\} - \log(|D_1 \cap D_2|)}{\log(|D|) - \min\{\log(D_1), \log(D_2)\}}$$

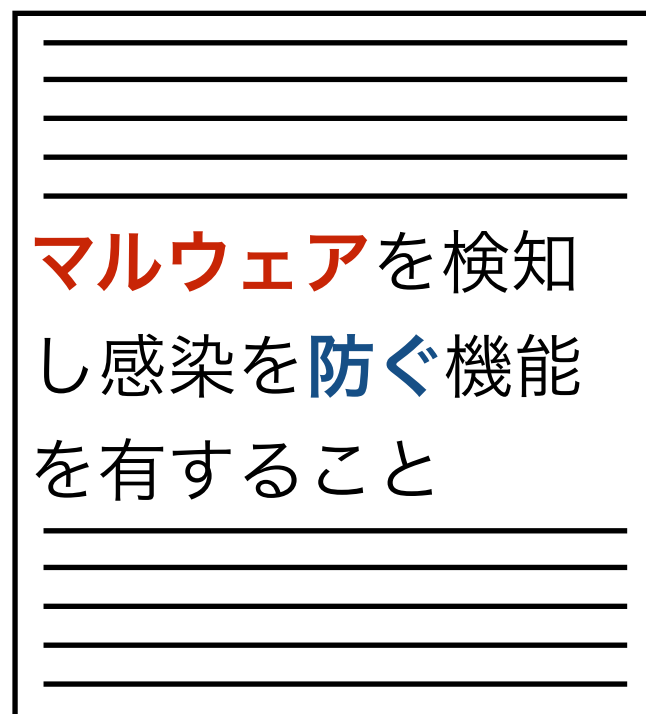
D : インデックスされているページの総数

D_1 : キーワード w_1 を含むページ数

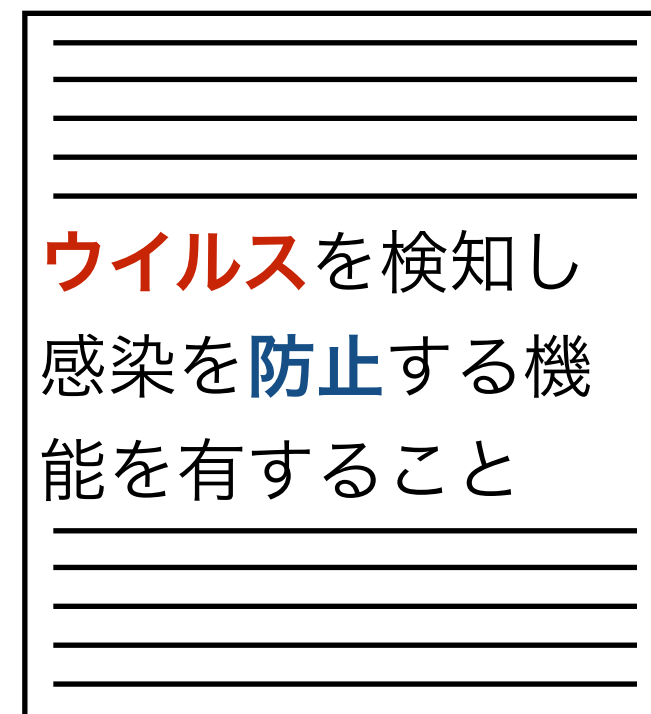
D_2 : キーワード w_2 を含むページ数

既存手法[5]の問題点

- ベンダーは様々な企業のシステム開発を担当しており、企業によって記述方法が異なる



企業Aの要件定義書



企業Bの要件定義書

- 類義語は同じ文書内に共起することが少なく、関連度を高く計算することができない

研究の目的とアプローチ

- 目的

- 類義語対応可能な非機能要件自動抽出手法を構築すること
 - 本研究ではセキュリティ要件に特化

- アプローチ

- セキュリティ要件専用の記述ルールを作成
 - 類義語に対応するためにWord2Vecを適用
 - 相関ルールマイニングを適用し記述ルールを抽出
- 要件定義書から記述を取得し記述ルールにマッチした記述を抽出

提案手法の概要

ステップ1：類義語の特定

1. 要件定義書から単語を抽出
2. Word2Vecを適用し単語のベクトル化

ステップ2：セキュリティ要件記述ルールを作成

1. 目視によるセキュリティ要件の特定
2. 相関ルールマイニングを用いた記述ルールの抽出

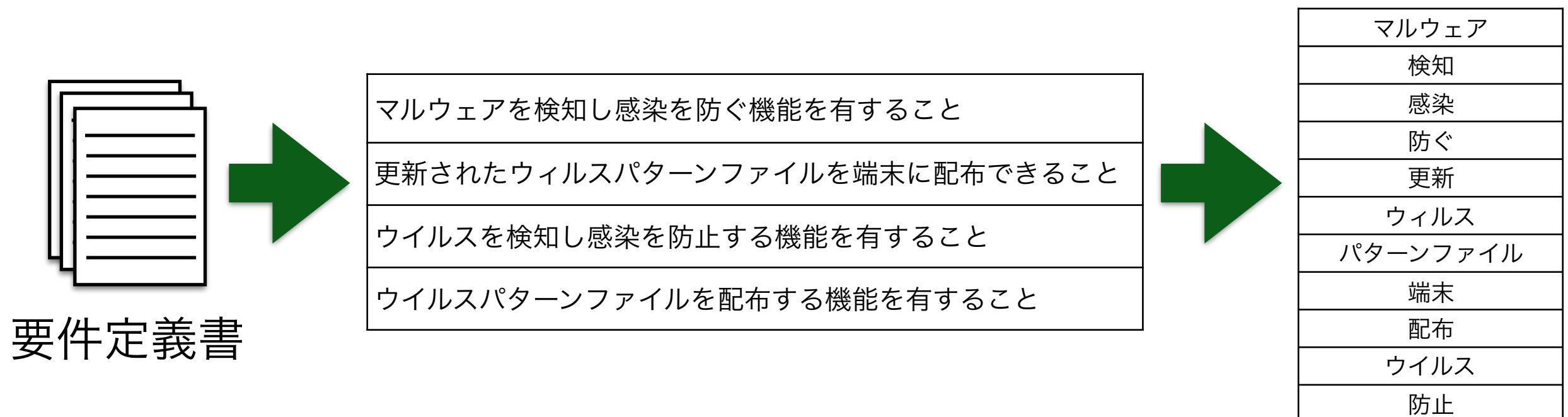
ステップ3：記述ルールを用いた記述の抽出

ステップ1

類義語の特定

1-1. 要件定義書から単語を抽出

- 要件定義書を1文または複文に分割
 - 1文ずつ分割
 - 文頭が接続詞の文は直前の文を修飾する文であるため、直前の文とマージ
- 形態素解析を行い単語に分割
 - ステミング処理を行い，出現頻度の低い単語を除去



1-2. Word2Vecを用いた単語のベクトル化

- Word2Vec

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

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

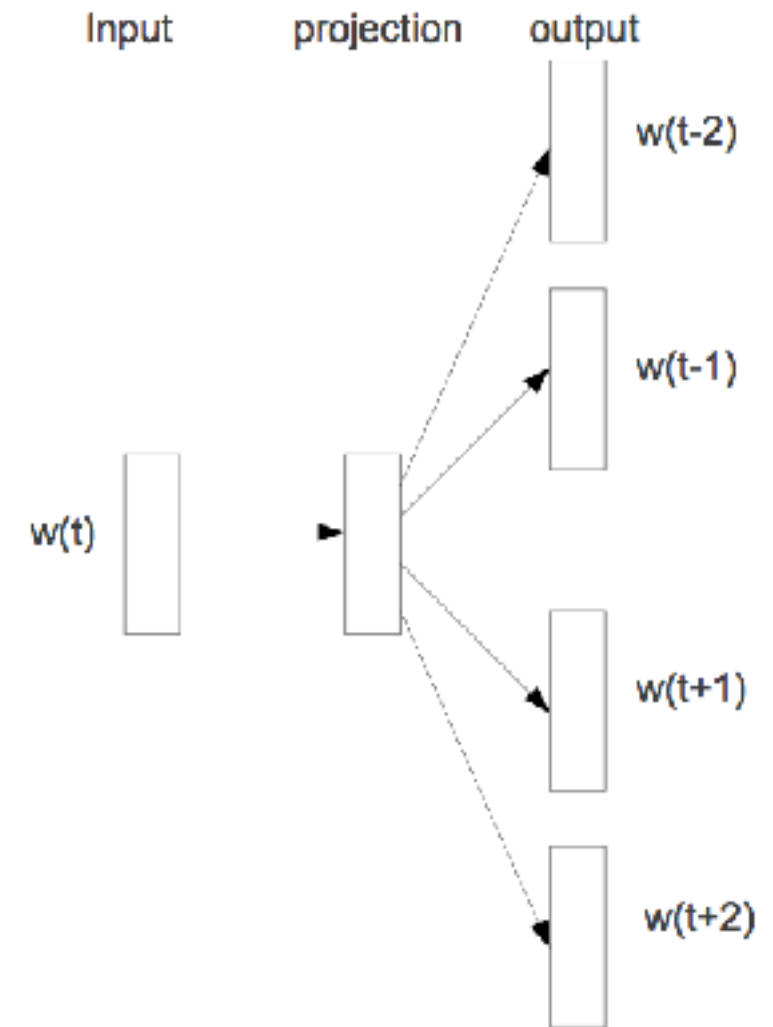
$$p(w_{t+j} | w_t) = \frac{\exp(v_{w_{t+j}}'^T v_{w_t})}{\sum_{w=1}^W \exp(v_w'^T v_{w_t})}$$

w_t : 単語

T : 学習する単語の総数

C : 学習する単語の周辺単語の数

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



Skip-gramの概要図

1-2. Word2Vecを用いた単語のベクトル化

.....

- 類義語を特定するために抽出した各単語にWord2Vecを適用しN次元ベクトルに変換
 - 次元数は任意で設定

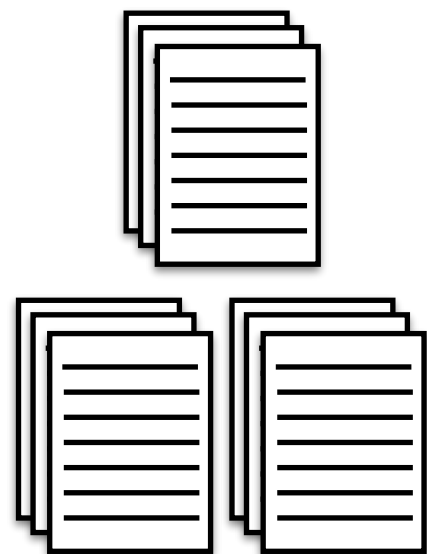
マルウェア	$[-0.6, 0.4, 0.3, \dots, 0.7]$
検知	$[0.2, -0.2, -0.8, \dots, 0.8]$
感染	$[0.7, 0.4, 0.5, \dots, -0.2]$
防ぐ	$[-0.7, 0.8, -0.1, \dots, 0.4]$
ウイルス	$[-0.3, 0.7, 0.1, \dots, 0.7]$
⋮	⋮
⋮	⋮
⋮	⋮

ステップ2

記述ルールの作成

2-1. 目視によるセキュリティ要件の特定

- セキュリティ要件の記述ルールを作成するために、要件定義書から該当する記述を特定する



要件定義書

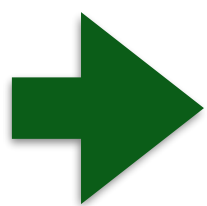
マルウェアを検知し感染を防ぐ機能を有すること	マルウェア対策の実施範囲
更新されたウィルスパターンファイルを端末に配布できること	マルウェア対策の実施範囲
ウイルスを検知し感染を防止する機能を有すること	マルウェア対策の実施範囲
パッチ、最新アップデートプログラム、バージョンアップ及びサポート切れ 情報等の更新、導入及び対応の対象は受注者が導入するすべてのハードウェア、ソフトウェアにおいて、OS、ミドルウェア、ハードウェア機器のファームウェアを対象とする。	セキュリティパッチの適用範囲
電子データの送受信を行う際には電子署名や暗号化を用いることで偽造等から 保護することが可能であること	伝送データの暗号化の有無
ネットワーク上のデータ通信の暗号化を実施することにより、盗聴・漏洩等の技術的な脅威に対し、システムの機密性を確保すること。	伝送データの暗号化の有無

2-2. 単語クラスタの記述ルールを作成

- セキュリティ要件の記述を満たす単語の共起関係を取得するために相関ルールマイニングを適用し，記述ルールを作成

例) マルウェア対策の実施範囲

	プログラム, ウイルス, ウィルス	パターン ファイル	端末	感染	更新	検知	配布	防止 防ぐ
1	1	0	0	1	0	1	0	1
2	1	1	1	0	1	0	1	0
3	1	0	0	1	0	1	0	1
4	1	1	0	0	0	0	1	0



1	[プログラム, ウイルス, ウィルス][検知][防止, 防ぐ]
2	[プログラム, ウイルス, ウィルス][パターンファイル][配布]

記述ルール

ステップ3

記述ルールを用いた記述の抽出

3-1. 記述ルールを用いた記述の抽出

- 要件定義書から記述ルールの単語クラスタを全て含んでいる記述を非機能要件の記述として抽出

マルウェアを検知し感染を防ぐ機能を有すること

ウイルスを検知し感染を防止する機能を有すること

ルール1を満たしており,
”マルウェア対策の実施範囲”
の記述として抽出

1	[プログラム, ウイルス, ウィルス][検知][防止, 防ぐ]
2	[プログラム, ウイルス, ウィルス][パターンファイル][配布]

本研究の分類対象

- 既存手法を適用した場合，開発者は分類結果を更に目視して分類する必要がある

大項目	小項目	メトリクス
セキュリティ	情報セキュリティに関する コンプライアンス	順守すべき社内規程、ルール、法令、 ガイドライン等の有無
	セキュリティリスク分析	リスク分析範囲
	セキュリティ診断	ネットワーク診断実施の有無
		Web診断実施の有無
		DB診断実施の有無
	セキュリティリスク の見直し	セキュリティリスク見直し頻度
		セキュリティリスクの見直し範囲
	セキュリティリスク対策の 見直し	運用開始後のリスク対応範囲
		リスク対策方針
	セキュリティパッチ適用	セキュリティパッチ適用範囲
		セキュリティパッチ適用方針
		セキュリティパッチ適用タイミング

本研究の分類対象

- 既存手法を適用した場合，開発者は分類結果を更に目視 既存手法の分類対象 がある 本研究の分類対象

大項目	小項目	メトリクス
セキュリティ	情報セキュリティに関する コンプライアンス	順守すべき社内規程、ルール、法令、 ガイドライン等の有無
	セキュリティリスク分析	リスク分析範囲
	セキュリティ診断	ネットワーク診断実施の有無
		Web診断実施の有無
		DB診断実施の有無
	セキュリティリスク の見直し	セキュリティリスク見直し頻度
		セキュリティリスクの見直し範囲
	セキュリティリスク対策の 見直し	運用開始後のリスク対応範囲
		リスク対策方針
	セキュリティパッチ適用	セキュリティパッチ適用範囲
		セキュリティパッチ適用方針
		セキュリティパッチ適用タイミング

現在の目標と進捗

- 目標

- PROMISE2018にフルペーパーを投稿

- 進捗

- 提案手法・既存手法の実装
- PROMISEデータセット[7]を用いて予備実験

[7] <http://openscience.us/repo/requirements/requirements-other/nfr.html>

予備実験

.....

- データセット

カテゴリ	件数
Availability	21
Fault Tolerance	10
Legal	13
Look and Feel	36
Maintenance	17
Operability	62
Performance	53
Scalability	21
Security	66
Usability	66
Functional Requirement	255
total	625

- 10 cross validationを実施, Precision・Recall・F-score・Accuracyで評価

Promiseデータセット (イメージ)

Pr	RequirementText string	class
1	'The system shall refresh the display every 60 seconds.'	PE
1	'The application shall match the color of the schema set forth by Department of Homeland Security'	LF
1	' If projected the data must be readable. On a 10x10 projection screen 90% of viewers must be able to read Event / Activity data from a viewing distance of 30'	US
1	' The product shall be available during normal business hours. As long as the user has access to the client PC the system will be available 99% of the time during the first six months of operation.'	A
1	' If projected the data must be understandable. On a 10x10 projection screen 90% of viewers must be able to determine that Events or Activities are occurring in current time from a viewing distance of 100'	US
1	'The product shall ensure that it can only be accessed by authorized users. The product will be able to distinguish between authorized and unauthorized users in all access attempts'	SE
1	'The product shall be intuitive and self-explanatory. : 90% of new users shall be able to start the display of Events or Activities within 90 minutes of using the product.'	US
1	'The product shall respond fast to keep up-to-date data in the display.'	PE
1	'The system shall have a MDI form that allows for the viewing of the graph and the data table.'	F
1	'The system shall display Events in a vertical table by time.'	F
1	'The system shall display the Events in a graph by time.'	F
1	'The system shall display Events or Activities.'	F
1	' The display shall have two regions: left 2/3 of the display is graphical right 1/3 of the display is a data	F

実験結果

カテゴリ	Precision	Recall	F-Score	Accuracy
Availability	0.22	0.79	0.33	0.72
Fault Tolerance	0.00	0.00	0.00	0.00
Legal	0.00	0.00	0.00	0.00
Look and Feel	0.06	0.10	0.07	0.23
Maintenance	0.12	0.43	0.17	0.33
Operability	0.22	0.64	0.21	0.33
Performance	0.81	0.44	0.55	0.85
Scalability	0.05	0.85	0.09	0.23
Security	0.52	0.38	0.41	0.79
Usability	0.14	0.94	0.24	0.30
Functional Requirement	0.00	0.00	0.00	0.00
total	0.24	0.45	0.22	0.39

Precision高 (0.5以上)

Recall高 (0.5以上)

- 予測できていない (0の値) NFRは、各文で独立した単語が使われて、相関ルールがない可能性ある

今後の課題

- 手法の改善
 - Word2Vecの学習
 - Wikipediaのデータで学習（データ量としては十分）
 - 各NFRごとに特徴的な単語のみを用いて相関ルールマイニングを適用
 - NFRは表現方法が異なるため，対象の単語を減らして少ない割合の記述ルールも抽出しやすくする
 - tf-idfを適用して，値の上位の単語を特徴語とする
- PROMISEデータのNFRをISO25010のNFRに改めて分類
 - ISO25010は詳細な分類をしている
- 評価実験

評価実験について

• 現在の実施方法

Req. ID	Requirement Text string	Class
1	'The system shall refresh the display every 60 seconds.'	FE
1	'The application shall match the color of the screens set forth by Department of Homeland Security.'	UF
1	'If projected, the data must be readable. On a 10x10 projection screen, 80% of viewers must be able to read Event / Activity data from a viewing distance of 30'.	US
1	'The product shall be available during normal business hours. As long as the user has access to the client PC, the system will be available 95% of the time during the first six months of operation.'	A
1	'If projected, the data must be understandable. On a 10x10 projection screen, 90% of viewers must be able to determine that Events or Activities are occurring in current time from a viewing distance of 100'.	US
1	'The product shall ensure that it can only be accessed by authorized users. The product will be able to distinguish between authorized and unauthorized users in all access attempts.'	SE
1	'The product shall be intuitive and self-explanatory. : 90% of new users shall be able to start the display of Events or Activities within 90 minutes of using the product.'	US
1	'The product shall respond fast to keep up-to-date data in the display.'	FE
1	'The system shall have a MDI form that allows for the viewing of the graph and the data table.'	F
1	'The system shall display Events in a vertical table by time.'	F
1	'The system shall display the Events in a graph by time.'	F
1	'The system shall display Events or Activities.'	F
1	'The display shall have two regions: left 2/3 of the display is graphical, right 1/3 of the display is a data table'	F

データセット

訓練データ（全体9割）

テストデータ（全体1割）

×10

- 訓練 or テストデータ内に存在しないNFRがある

- ▶ 先程見せた結果は，テストデータ内にNFRがない場合の結果は覗いた平均値をとっている（8回分で割っていたり）
- ▶ たぶん，というかほぼほぼよくない．．．

考えている実験方法

1. どちらにも全NFRが含まれる用にして10 cross validation
2. leave one out
 - 「全部の中から1個だけデータをとってきてテストデータにして残りを訓練データにする」 × データ数回やる方法
 - NFRごとに実施