

春休み成果報告

60180132

宮崎智己

取り組んでいること・目標

- 取り組んでいること
 - 非機能要件自動抽出手法の構築
 - ▶ 要件定義書や調達仕様書に書かれている非機能要件（NFR）を自動抽出する
- 目標
 - 先行研究と同程度の精度で粒度の細かい分類を
 - ▶ “セキュリティ”と分類するのではなく“セキュリティ・マルウェアに関する項目”と分類

春休みに行ったこと

1. 仕様書とWikiをそれぞれ学習データにしたWord2Vecモデルの作成
 - 単語の意味を考慮できるように
2. 抽出方法のプロトタイプ版の実装
 - 先行研究を参考に実装
 - 改善の余地しかない
3. NFR記述文の単語分布の調査
 - NFR辞書を教師なし学習でできるか判断するため

3.単語の分布調査

- 目視したNFRの文章の単語の分布を調査
 - 先行研究よりも細かい粒度で分類を行うためには, NFRごとに記述文の単語分布が異なる必要がある
 - 性能性, セキュリティ性ごとに単語の分布を調査
 - 単語ごとにzスコア正規化を行っている

性能性	利用者	レスポンス	...
ユーザ数	2.48	-5.11	• •
応答時間	-1.22	4.55	• •
• •	• •	• •	• •

単語の分布調査

- 特徴語（目視の際に判断基準となった単語）は比較的多く出現している？
 - ただし、最も多いわけではない．．．

	遵守すべきルールの有無		セキュリティパッチ適用範囲	
	単語	スコア	単語	スコア
1	情報セキュリティ	5.363	ミドルウェア	4.742
2	基準	5.293	共通	4.649
3	財務省	5.292	OS	4.303
4	厚生労働省	5.170	開始	4.047
5	講じる	5.077	担当	4.023
6	書	4.675	業者	3.983
7	行政	4.520	セキュリティホール	3.932
8	提案	4.047	セキュリティパッチ	3.925

今後の予定

- 手法の改良
 - このままでは，先行研究をなぞっただけ
 - 抽出方法、辞書の作成方法を提案
- 先行研究の手法の実装
 - 比較実験用
 - ライブラリではないが参考にできるスクリプトを発見
- 正解のデータセット作成
 - SRAと一緒に5月頃？に行う

抽出方法の改善

- 特徴語と類似した単語を全て含む単語を抽出

利用者
数



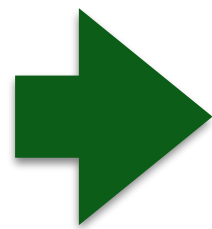
ユーザ数は1000人である



ユーザは職員、または国民である

- 文は自然言語で構成されるため、必ずしも特徴語を全て含むとは限らない

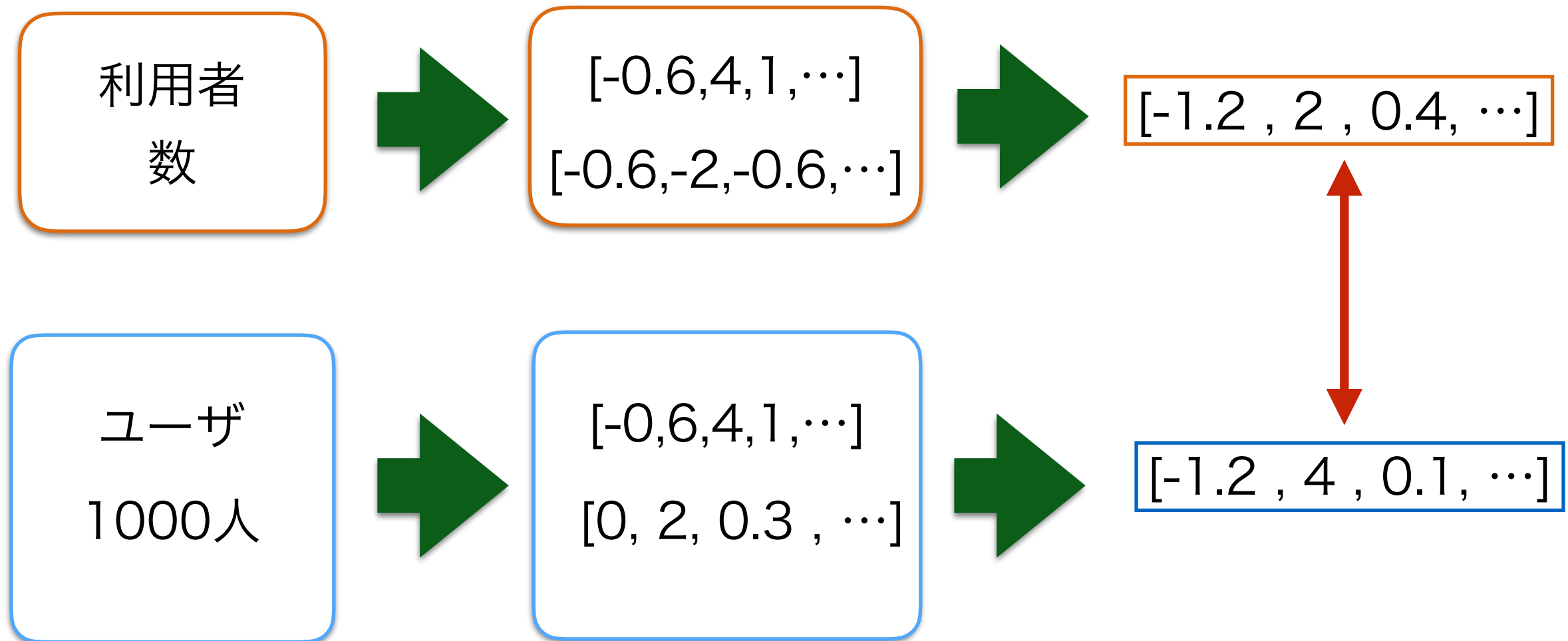
本システムは本職員1000人が利用する



ユーザ数に関する記述であるが、上で定義している特徴語を含まないため、抽出できない

抽出方法の改善

- ・ 単語ではなく文をベクトル化して類似度を比較する
 - ・ 文をベクトル化することで、文の意味を考慮する
 - ・ word2vecの改良版であるDoc2vecを用いる



辞書の作成

- 文章をDoc2Vecを用いてベクトル化し，文の類似度を計算する
- 類似度の高い文をクラスタリングして，作成したクラスタを辞書とする
 - 文に必要な形態素を明らかにする（名詞のみでいいのか，名詞＋動詞なのか）

論文執筆予定

- SESもしくはFOSEに投稿する
 - 大ゼミ，修論のためにも必ず投稿
 - SESに投稿する前提で動く（大ゼミの発表も、、、）

	4月 前	4月 中	4月 後	5月 前	5月 中	5月 後	6月 前	6月 中	6月 後	7月 前	7月 中
手法の改良 案を考える											
実装											
実験											
論文執筆											
大ゼミ 発表？											

以降, 研究発表スライド

研究の背景

- 近年のシステム開発では、委託開発が主流
- ベンダ側は開発するシステムに対する顧客の要件を正確に把握する必要がある
 - RFP（調達仕様書）や要件定義書から必要な要件を把握

要件	概要
機能要件	システムが満たすべき 機能に関する要件のこと
非機能要件	機能以外の要件のこと システムの性能やセキュリティなど

非機能要件（NFR） [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]非機能要件グレード, <https://www.ipa.go.jp/sec/softwareengineering/reports/20100416.html>

NFR特定の課題

- NFRは非常に多くの項目が対象
- RFPや要件定義書は多くの自然言語で記述
 - 多いものでは200ページ以上にも

1. 調達案件の概要に関する事項

(1) 調達件名

労働基準行政システムの業務改善に係るアプリケーションプログラム改修業務一式

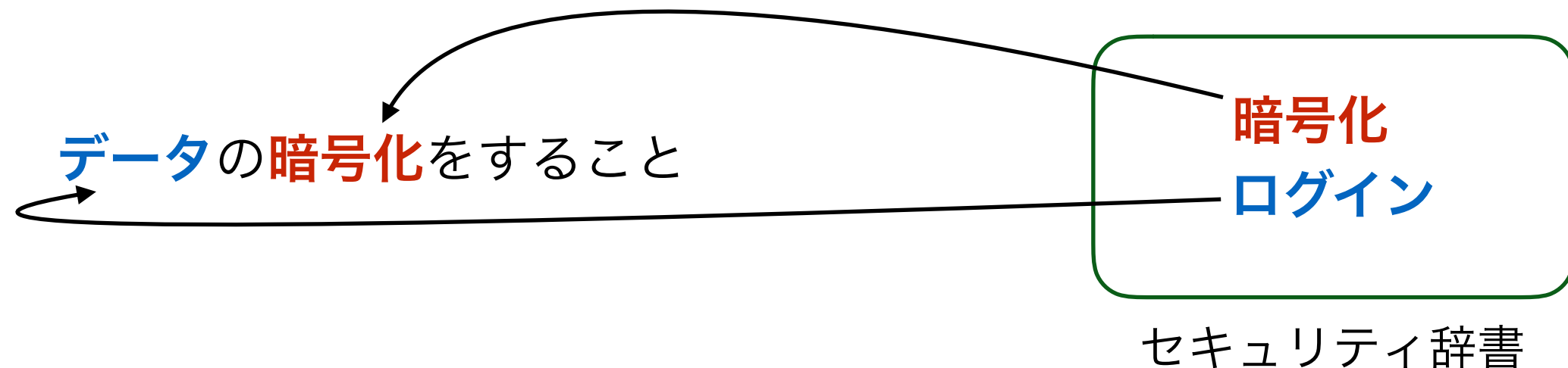
**RFPや要件定義書に含まれている
非機能要件を特定することは非常にコストがかかる**

また、第Ⅱ期開発分（業務処理の集中化等）として平成22年12月に基準システムが稼働し、平成23年5月には労災システムの第Ⅱ期開発分が稼働した。

上記の各種システムの稼働後に、保守維持等を目的として、労働基準行政情報システム・労災行政情報管理システム（以下「基準・労災システム」という。）のハードウェア・ソフトウェア（以下「ハードウェア等」という。）の更改を行って、基準・労災システムが平成25年5月から稼働している。

先行研究[2]

- RFPに含まれるNFRを自動で抽出する手法
 - 各NFRごとに辞書を作成
 - 辞書の単語に似ている単語を含んでいる文を，そのNFRの文として分類



データの暗号化をすること

先行研究[]

- 階層クラスタリングを用いて各NFRの辞書を作成
 - 単語の関連度によるクラスタリング
 - 関連度計算にはNGDを利用

**分類する項目の粒度が大きく，利用者は改めて
出力結果を手作業で分類する必要がある**

研究の目的

**RFP・要件定義書に含まれる非機能要件を
自動で抽出する手法の構築**

非機能要件抽出手順の概要

0. RFP・要件定義書から文を取得する

1. 文を形態素解析・名詞を取得

2. 1で得た単語と各NFRの特徴語との類似度を算出

- NFRの特徴語：各NFRを記述する際に必要となる単語

3. NFRの特徴語と一致・類似した単語を全て含んでいる場合、NFRの文として抽出する

自然言語処理の課題

- 同じ意味を指す文が異なる単語で構成されることがある
 - 通常は, 「単語」を「文字列」として認識するため, 同じ意味を持つ単語を全く異なる意味として扱ってしまう

「ユーザ」と「利用者」

「適用」と「充てる」

同じ意味を指すが
異なる単語として処理

- 単語の意味を考慮できるようにWord2Vecを用いる

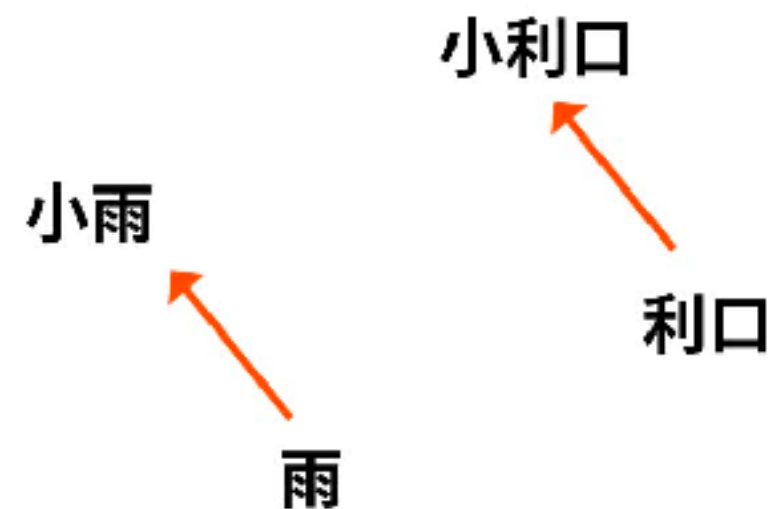
Word2vec

.....

- Word2vec

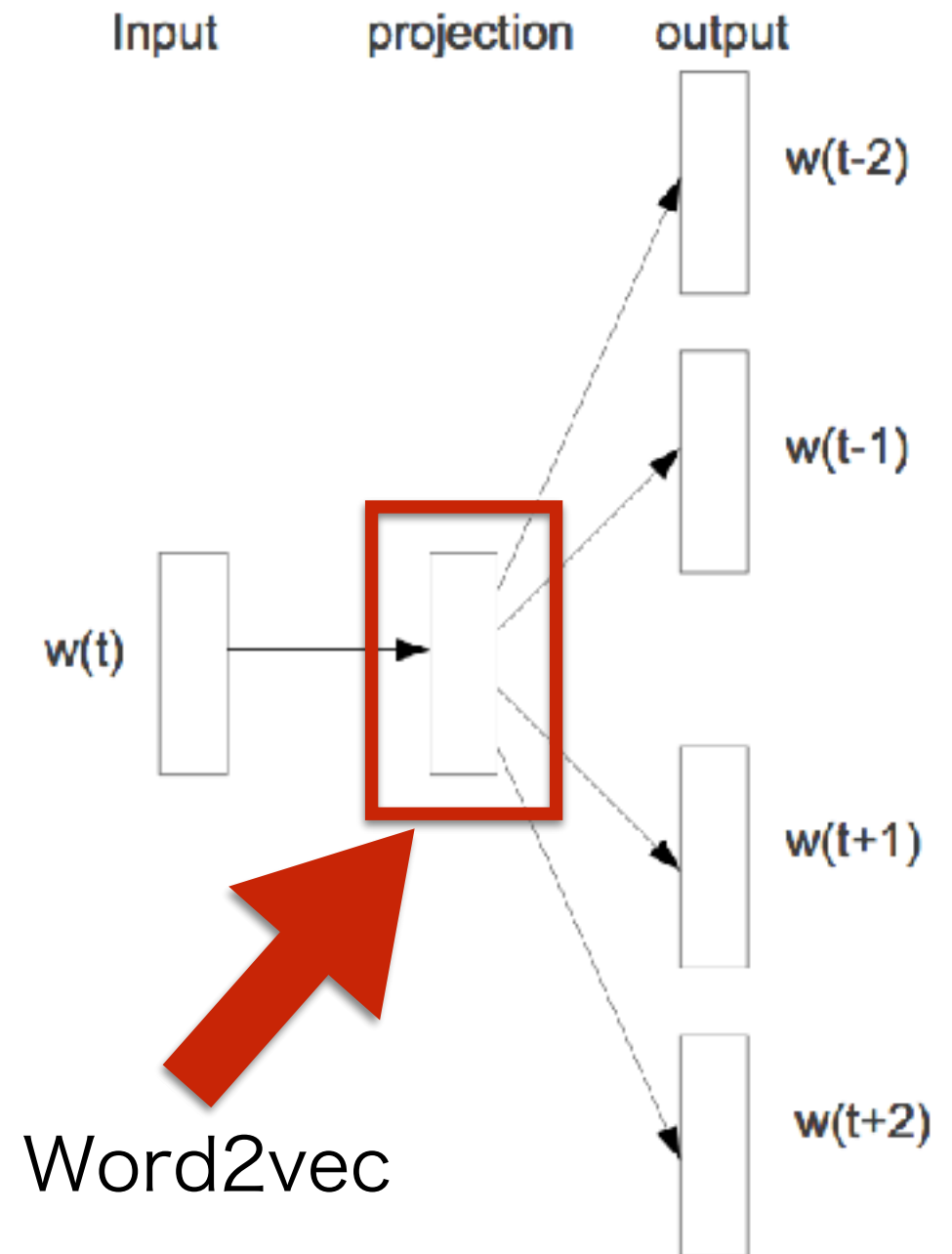
- 単語をm次元のベクトルで表現
- 単語をベクトル化することで、単語の**意味**を考慮するよ
うに振る舞うことが可能
- 単語の加減算・類似度の算出も可能

	イチロー	プログラマー	パリ
怠惰さ	0.01	0.82	0.34
スポーツ	0.99	0.14	0.6
土地	0.04	0.02	0.97
	⋮	⋮	⋮



Word2vec

- ある単語を入力したときに、期待する単語が出力されるように学習したベクトル
 - 「私は学生です」という文があった場合、
 - 「私」を入力したら次に「は」を出力するように学習



Word2Vecの学習データ

1. Wikipediaの記事データ

- 一般的な単語表現に対して対応できる
- データ数も多く，容易に取得可能

2. RFP・要件定義書などのデータ

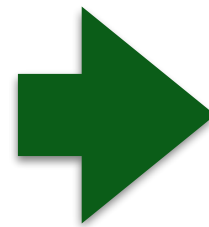
- 専門的な表現，プロジェクトの固有表現に対応できる
- Wikipediaと比べてデータ数が少ない？
 - ▶ 省庁や地方自治体などのRFP・要件定義書59件
 - ▶ 今後データの追加をする必要がある

1.形態素解析

- MeCabを用いて形態素解析を行い名詞のみを抽出
 - NFRの記述文の有無を確認するため
 - 辞書には「mecab-ipadic-NEologd」を利用
 - 通常の辞書とは異なり新語・固有表現を扱うことができる

(例) 利用者数に関する記述文

ユーザ数は1000人である



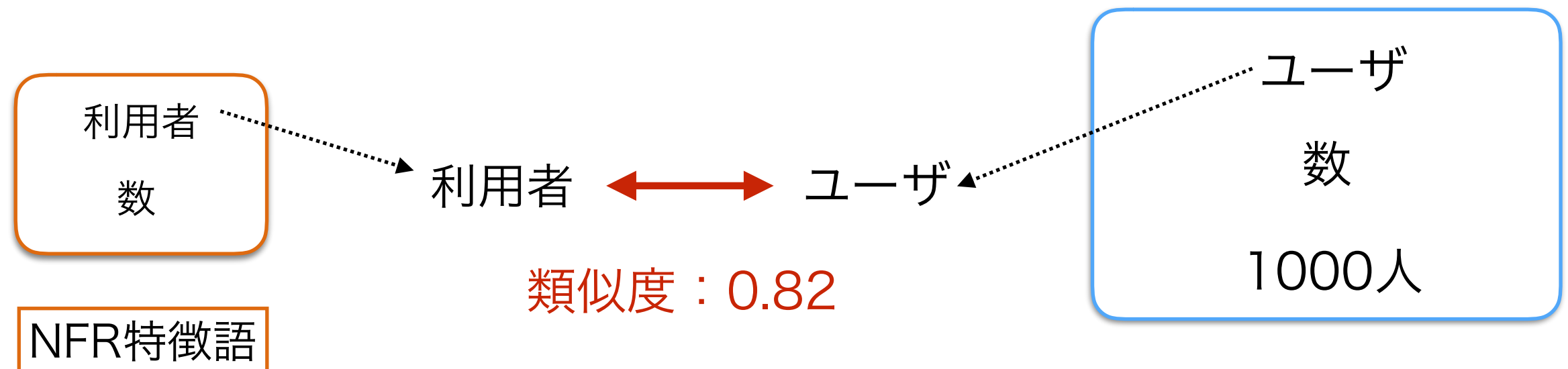
ユーザ

数

1000人

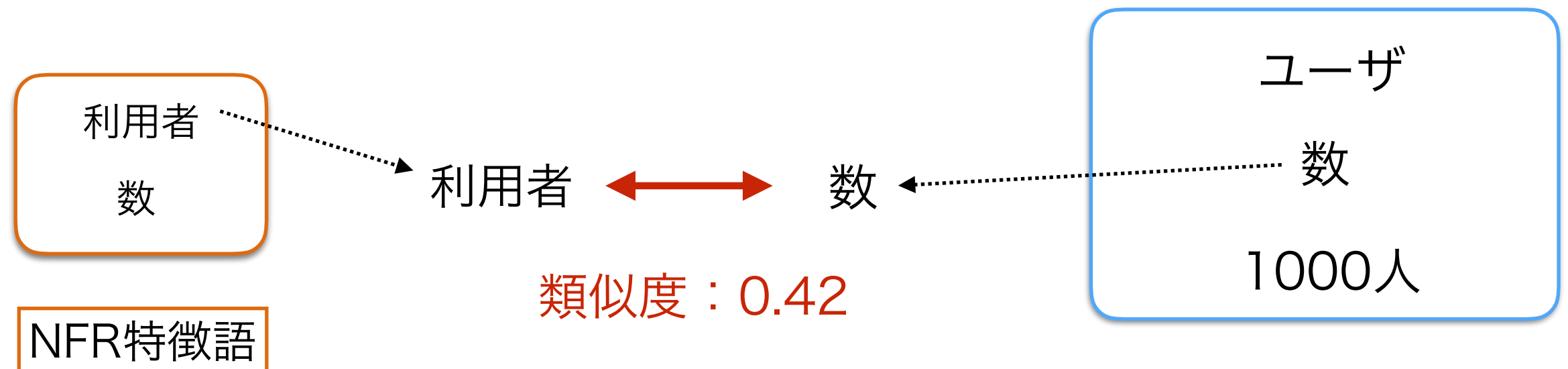
2.NFR特徴語との類似度を計算

- 1で得た単語とNFRの特徴語の類似度を計算
 - NFRの特徴語：各NFRを記述する際に必要となる単語
 - 目視結果から主観で設定
 - 例) 「パッチの適用」に関する特徴語：パッチ, 適用
 - 特徴語と文字列は異なるが同じ意味を持つ単語を特定するため



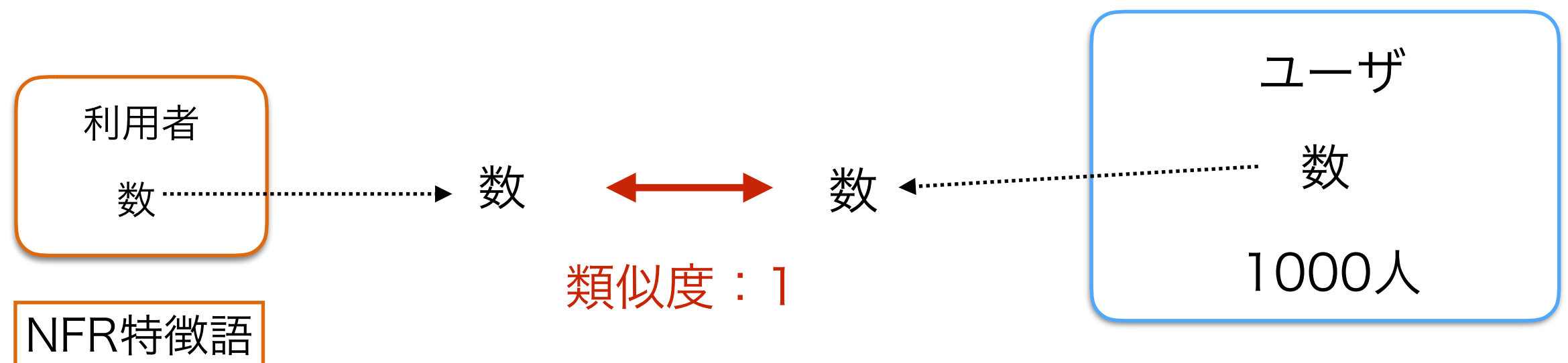
2.NFR特徴語との類似度を計算

- 1で得た単語とNFRの特徴語の類似度を計算
 - NFRの特徴語：各NFRを記述する際に必要となる単語
 - 目視結果から主観で設定
 - 例) 「パッチの適用」に関する特徴語：パッチ, 適用
 - 特徴語と文字列は異なるが同じ意味を持つ単語を特定するため



2.NFR特徴語との類似度を計算

- 1で得た単語とNFRの特徴語の類似度を計算
 - NFRの特徴語：各NFRを記述する際に必要となる単語
 - 目視結果から主観で設定
 - 例) 「パッチの適用」に関する特徴語：パッチ, 適用
 - 特徴語と文字列は異なるが同じ意味を持つ単語を特定するため



3.記述文の抽出

- 全ての特徴語と一致または類似した単語を含んでいる記述文をそのNFRを示す記述文として抽出する
 - NFRの項目が多く， 分類条件を細かく設定するため

利用者
数



ユーザ数は1000人である



ユーザは職員， または国民である

補足資料

前回のMTGでの懸念事項

- 要件定義書の表部分をテキスト化すると文章が崩れる
- pdfからテキスト化すると抽出できない文字がある場合がある
 - ・ セキュリティパッチ、OSアップデート、ファームウェアアップデート等の適用を多量の機器に対して迅速に実施できる方式とすること
 - 『セキュリティパッチ、OSアップデート、ファームウェアアップデート等の適用を多量の機器に対して迅速に実施できる方式とすること』

前処理

- pdf形式のRFPをAdobe Readerを使用してWordに変換
 - 表の構造を利用可能
- Unicode正規化
- 文一つ一つを改行で区切る
 - 箇条書きには句点がない場合があるため
 - ただし、接続詞・副詞始まりの文は直前の文と関連が高いためつなげて”一文”とする

Word2vecモデルの比較

(例) 「パッチ」の類似度上位5単語

Wikipedia

	単語	類似度
1	バグ	0.72
2	ファームウェア	0.710
3	プラグイン	0.706
4	ドライバ	0.683
5	ソースコード	0.672

RFP

	単語	類似度
1	たり	0.780
2	コピー	0.704
3	適用	0.674
4	ながら	0.637
5	セキュリティ パッチ	0.620

予備実験

- 実際にNFRに関する記述を抽出できるか実験
 - セキュリティパッチの適用・伝送データの暗号化について実施
- 対象ドキュメント
 - RFP・要件定義書合計10件
 - ※特徴語の選出には利用していない
- パラメータの設定
 - Word2vecの次元数：200次元
 - 類似度の閾値：0.6

結果：セキュリティパッチ適用

.....

モデル	抽出した文数	正解数
Wiki	22件	8件
RFP	11件	6件

- 抽出した文の半分がセキュリティパッチ適用に関する記述
 - RFPモデルの方は、「パッチ」を「セキュリティパッチ」の類似単語と判断

抽出した文の例

- Wikiモデル

抽出した文

(エ)製品サポート期間の確認本情報システムの構築等又は運用・保守・点検の際に導入する製品(ソフトウェア及びハードウェア)については、本情報システムのライフサイクル(システム利用期間の終了まで)におけるサポート(部品、セキュリティパッチ、不具合への対応としてのパッチの提供等)が継続される製品を導入すること。

2 ハードウェア・ソフトウェアの保守要件 ア 平常時対応導入したハードウェア及びソフトウェアの開発元・販売元からサポートを確実に受けられる状態を確保し、セキュリティパッチ、最新アップデートプログラム、バージョンアッププログラム及びサポート切れ情報を取得し、更新情報を厚生労働省に提供すること。

- 「導入」と「適用」を類似した単語として判定

抽出した文の例

- RFPモデル

抽出した文

OS等のパッチ

パターンファイルの更新については、ソフトウェアベンダー等において、パターンファイルが公開された時点で、迅速に本システムに適用できる仕組みを構築すること。また、利用者及び担当職員の作業負担のない方法を実現すること。

- 「パッチ」と「適用」，「ソフトウェアベンダー」と「セキュリティパッチ」をそれぞれ類似していると判定

結果：データ暗号化

.....

モデル	抽出した文数	正解数
Wiki	13件	3件
RFP	0件	0件

- 抽出ができていない
 - 特徴語の選出に問題があり？

抽出した文の例

.....

- Wikiモデル

抽出した文
インターネット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。
利用者のパスワード等の情報を暗号化して通信する機能を設けること。
(ウ) 暗号化機能重要情報等の秘匿を保持することで重要情報への不正アクセス、改ざん及び外部への持ち出しができないよう、暗号化について検討し、必要な機能を設けること。

- 「伝送」と「通信」を類似と判定できている反面，「伝送」と「暗号」も類似と判定

閾値の設定

- 0.8に設定した場合
 - 「セキュリティパッチの適用」については両モデルとも3件
 - 特徴語と完全一致した単語を含む記述文のみしか抽出できていない
- 類似度の閾値の設定を正しくする必要がある
 - 低すぎると全く使用目的が異なる単語を「類似した」単語として判別してしまう