E-023

# 業務改善に向けたダークデータ活用技術
# Dark Data Utilization for Improvements of Business Operations

照屋　絵理　　愛甲　和秀　　竹内 理
Eri Teruya　　Aikoh Kazuhide　Tadashi Takeuchi

## 1. はじめに

分析技術が進歩し，企業が保持しているビッグデータを活用して事業の競争優位性を創り出した事例が増えるにつれ，ダークデータ（収集・蓄積はされているが，活用されていなかったデータ）を活用／分析する技術への注目が高まりつつある[1]。ダークデータの多くはテキストデータや画像データなどの非構造データである。これらの非構造データを活用する際に鍵となる技術が，非構造データを構造データ化し，そこから有用な情報（知識）を抽出する技術である。特に企業が社内に保有する社内文書（社内公式文書，障害報告文書などのテキストデータ）には多数の知識が蓄積されていると言われており，その構造データ化技術に注目が集まっている[2]。テキストデータを構造化する技術の一つとして，分析に必要な固有表現のみを教師有り学習で抽出する Named Entity Recognition(NER)[3]がある。

しかし，一般に NER では学習に必要な教師データの準備に工数がかかるという問題があった。そこで我々は，高精度 NER を実現する Smart Dictionary 基盤を開発した。本基盤では， distant supervision による機械学習を行う I-NER (Interactive Named Entity Recognizer)と，抽出結果に対してユーザのフィードバックを受けることで，NER の精度を高めるフィードバック機能を備える。また，skip-gram[4]を活用した関連語/類義語抽出機能を持つ L-HRE (Label-based Hidden Relation Extractor)と連携することで，テキストデータを構造データ化することも可能にしている[5]。本研究では，Smart Dictionary 基盤を用いて実際の社内文書を用い，その精度と辞書整備工数の評価を行った。特に，フィードバック機能の評価を行った。

以降では，2 章で想定ユースケース，3 章で Smart Dictionary 基盤の概要，4 章で評価，5 章でまとめを述べる。

## 2. 想定ユースケース

鉄道等の社会インフラシステムは長期利用やそれに伴う頻繁な改修等によりシステムが複雑化しており，システム障害の低減が課題となっている。障害低減のため，インフラシステム製造企業では設計レビューを行っている。障害内容を分析すると多くが再発障害であり，根本原因や動機的要因が同一なため，新規システム設計時に過去に発生した障害の知見から新規システムに起こりうる障害を予め洗い出し，障害防止策の設計反映や不良の作りこみを排除することで障害が発生しにくい製品の設計開発が可能である。

しかし，設計レビューを行う設計者と障害報告書を作成する保守員とでは部署が分かれていて業務が分断されている。このため，設計者が過去の障害報告書を設計レビューに 2 次利用する際，保守員部署特有の言い回しがあるため，社内で類義語や表記ゆれが発生し検索キーワードを間違える，また，障害原因の因果関係や製品の構成要素などその部署しか知らない暗黙的な単語間のつながり（関連性）があるため検索キーワードが不足するなど，膨大な障害報告書から求める障害報告書の検索が困難となるという問題が発生した。そこで，テキストデータから予め構造データを抽出し，本構造データをもとに装置名や異常動作名と関連する処理名やデータ名などの検索キーワードを自動追加できるようになっていれば，障害報告書の検索が可能になる。

## 3. Smart Dictionary 基盤

前記問題を解決するため，我々は Smart Dictionary 基盤を開発した。本基盤は，図 1 に示すシステムアーキテクチャを持ち，NER を行う I-NER と省工数で NER の精度を高めるフィードバック機能を備える。

### 3.1 I-NER

I-NER はユーザが入力する少量のドメイン知識情報から作成した初期正解（不正解）辞書（初期正解（不正解）辞書：少量の抽出対象（抽出対象でない）の Named Entity(NE)の単語リスト）を元に半教師あり学習を行い，NER を実現する。I-NER は初期正解（不正解）辞書内の単語から頻出する特徴量の要素を学習する。特徴量とは単語の最終単語や左右の単語など文書中のそれぞれの単語が持つ単語自身や文章上の性質である。また，特徴量の要素の例として，図 1 に示すように「BBB 装置が起動しませんでした。」という文章内の「BBB 装置」という単語の場合，最終単語という特徴量に対し「装置」，右単語という特徴量に対し「が起動」などがある。また，学習によりそれぞれの特徴量の要素が NE を表す特徴として良い，悪い，良くも悪くもない，を判断する。さらに，学習により判断された特徴量の要素の良し悪しから推論し，NE らしさを表す推論確率を 0 から 1 連続値で出力する。設定した閾値より高確率の単語を NE である，閾値より低確率の単語を NE ではないと判断する。

### 3.2 フィードバック機能

I-NER にて出力された NE 辞書の一部の単語をユーザに提示し，ユーザに NE の正誤のフィードバック情報を入力してもらう。その内容を初期辞書に加え再学習することで，低コストで精度の向上を図る。下記にフィードバック機能の処理概要を記載する。

フィードバック機能の処理概要

1) I-NER により出力された A-4 の NE 辞書のうち，B にて一部の単語をフィードバック GUI を通しユーザに提示する。
2) 提示した結果のラベル値が正しく付与されているかどうかの正誤情報のフィードバックをフィードバック GUI を通してユーザから受け取る。
3) 正誤情報をもとに A-3 のラベル値を変更し再学習を行い確率モデルを更新する。

(株) 日立製作所　Hitachi, Ltd.

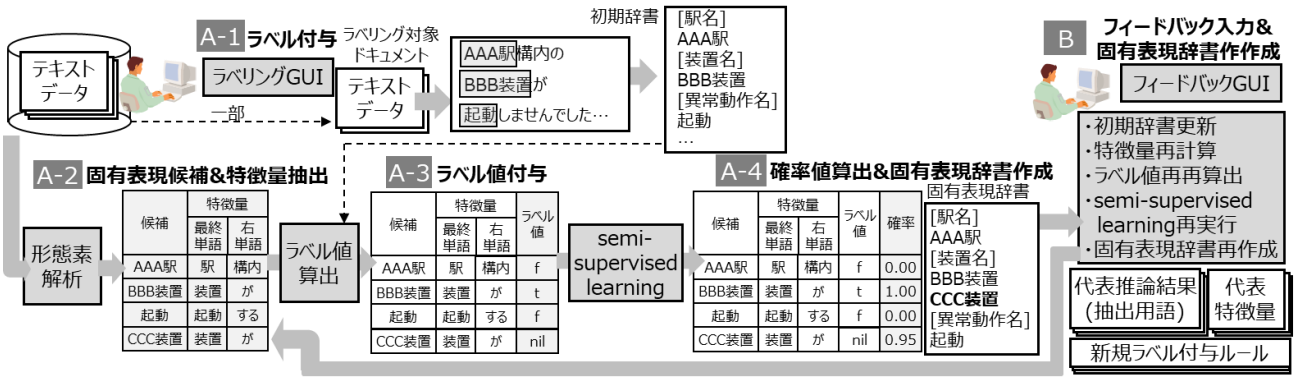


図 1　Smart Dictionary の実現方式

1)における単語の提示方法について，高精度の辞書を実現するため，フィードバックにより高い改善効果が見込まれる，すなわち「確率モデルに大きく寄与する単語」を提示する。また，「少ないフィードバック回数」を実現するために「誤判定である確率の高い単語」を提示する。さらに，双方を加味し「確率モデルに大きく寄与し，かつ抽出結果が誤判定である可能性の高い単語」を提示する。

#### 3.2.1 確率モデルに大きく寄与する単語の提示

確率モデルに大きく寄与する単語の提示方法として次の 2 パターンを用いた。

(A)：ドキュメント中に頻出する単語を提示する。頻出する単語は，単語が出現する場所毎に特徴量の要素を持っているため，多くの特徴量を持っている。このような単語の正誤情報をフィードバックしてもらった場合，より多くの特徴量の要素の良し悪しが判断されるため，確率モデルが大きくエンハンスされる。

(B)：ドキュメント中の単語が持つ特徴量の要素のうち，多くの単語が持つ特徴量の要素（例えば図 1 の最終単語：装置），すなわち頻出する特徴量の要素を持つ単語を提示する。このような単語の正誤情報をフィードバックしてもらった場合，より多くの特徴量の要素の良し悪しが判断されるため，確率モデルが大きくエンハンスされる。

#### 3.2.2 抽出結果が誤判定である可能性の高い単語の提示

抽出精度を向上させるには，(1)NE でないのにも関わらず確度が閾値以上であった単語を排除，および(2)NE であるのにも関わらず確度が閾値未満であった単語を抽出する必要がある。よって，本研究では次の 2 パターンを用いた。

(C)-(1)：閾値以上の確率を持つ単語のうち，単語が持つ特徴量の要素と初期正解（不正解）辞書内の単語が持つ特徴量の要素の overlap が小さい（大きい）単語を提示する。

(C)-(2)：閾値未満の確率を持つ単語のうち，単語が持つ特徴量の要素と初期正解（不正解）辞書内の単語が持つ特徴量の要素の overlap が大きい（小さい）単語を提示する。

#### 3.2.3 確率モデルに大きく寄与し，かつ抽出結果が誤判定である可能性の高い単語の提示

(A)-(C)を全て加味し，次の 2 パターンの単語を用いた。

(D)-(1)：(A),(B),(C)-(1)を全て加味した単語を提示する。

(D)-(2)：(A),(B),(C)-(2)を全て加味した単語を提示する。

## 4. 評価

### 4.1 評価方法

本研究では実業務で作成した鉄道の障害報告書 135 件から Smart Dictionary 基盤の I-NER にて装置名を抽出し，さらにフィードバック方法①から③を用いたフィードバック機能を使用して 20 単語をユーザに提示し，フィードバックを受けることによる精度の改善効果を評価した。

### 4.2 評価結果

I-NER の結果，図 2 に示す通り precision:0.66, recall:0.69, F-measure:0.62 となった。また，フィードバック機能により 20 単語をユーザに提示し，正誤情報の入力を受け付けたところ，精度が最大で precision:0.88, recall:0.67, F-measure:0.76 となり，F-measure が 0.14 ポイント向上した。

## 5. おわりに

本研究では，企業の社内文書に含まれるテキストデータを構造データ化し有用な情報（知識）を抽出する際に必要となる，構造データ化基盤 Smart Dictionary 基盤を開発した。本基盤は，I-NER およびフィードバック機能を備える。本基盤を用いて障害報告書から装置名を抽出する精度を評価した。この結果，フィードバック機能により F-measure が 0.14 ポイント向上した。今後，他の企業内文書に本基盤を適用することにより，本基盤の有効性を検証する。

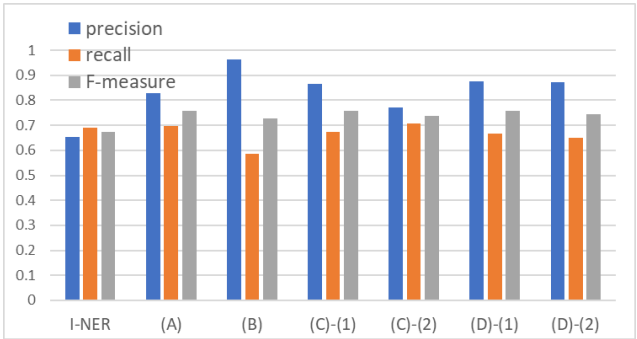


図 2　NER 精度の比較

### 参考文献

[1] 1) Njeru Mwiti Kevin,et al., “Dark data: Business Analytical tools and Facilities for illuminating dark data”, Scientific Research Journal, Volume IV, Issue IV (2016)
[2] Luiz Gomes,et.al,: Information Extraction in the Business Intelligence Context, (2010)
[3] Vikas Yadav,et al.: A Survey on Recent Advances in Named Entity Recognition from Deep Learning Models, Proceedings of the 27th, CICLing, pp2145–2158 (2018).
[4] Tomas Mikolov,et al: Distributed Representations of Words and Phrases and their Compositionality,
[5] 竹内 理, “社内文書から固有表現を抽出する Smart Dictionary 基盤の開発”, デジタルプラクティス, in preparation.