

## ユーザフィードバックによる固有表現抽出の精度改善 Improving Named Entity Recognition using User Feedback

加藤 大羽<sup>†</sup> 田中 美智子<sup>†</sup>  
Daiba Kato Michiko Tanaka

### 1. はじめに

近年、ダークデータ（企業内に眠っていて、二次利用されていない非構造データ）を生産性向上やコスト削減等に活用する取り組みが注目されてきている[1]。テキスト系ダークデータをビジネスに活用するためには、フリーフォーマットで記載されている文書からビジネスに有用なデータを抽出し、従来の IT システムで処理しやすいテーブル形式に変換する必要があり、膨大な時間をかけて行っている[2]。この問題解決に向け、我々は膨大な文書からビジネスに必要なデータを半自動的に抽出する Smart Dictionary 基盤を開発した。Smart Dictionary 基盤は、ドメイン知識を持つユーザ（SME, Subject Matter Expert）が入力する限られた量のドメイン知識情報（初期辞書）を準備することで、省工数で固有表現抽出（NER, Named Entity Recognition）を実現する[3]。

しかしながら、固有表現抽出技術によってビジネスに有用なデータを高精度に抽出するためには、抽出精度の改善作業に取り組むことが必要となる。その精度改善作業にはドメイン知識を持つ SME と、特徴量チューニングを行うエンジニアが連携して作業に当たらねばならず、人手作業の負担が多く工数削減が課題であった。そこで我々は、Smart Dictionary 基盤に SME 単独でも精度改善を可能とするユーザフィードバック機能を追加することで、人手作業の工数削減をしつつ NER 精度を改善する機能を開発した。まず Smart Dictionary 基盤が抽出した結果のうち、いくつかの単語について SME が正誤判定を行う。その正誤判定情報に基づいて再学習を行うことで SME 単独での精度改善を実現する。このユーザフィードバック機能による抽出精度の改善効果および工数削減の見積りは、実際の社内文書からの情報抽出にて評価を行った。

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

企業が社内に保有する大量のテキストデータの構造データ化には高精度な辞書が必須であるが、既存技術では辞書整備に工数がかかり、業務効率化のコストメリットを得られなかった。そこで、テキストデータから省工数で構造データが抽出できるようになれば、人名や顧客先といった情報と課題名や製品名などを関連付けてデータ整理することが容易になり、類似事例等の迅速な検索が可能になる。

### 3. Smart Dictionary 基盤

前記問題を解決するため、我々は Smart Dictionary 基盤におけるユーザフィードバック機能を開発した。Smart Dictionary 基盤は、図 1 に示すシステムアーキテクチャを持ち、固有表現抽出を行う I-NER（Interactive Named Entity Recognition）と省工数で抽出精度を高めるユーザフィードバック機能を備える。

#### 3.1 I-NER

I-NER は図 1 に示すように、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辞書を元に半教師あり学習を行い、固有表現抽出を実現する。固有表現抽出とは、文中から単語抽出し、あらかじめ定義されたカテゴリ（人名・組織名・地名・日付・数量など）へ分類することで機械が文書構造を理解する技術である。I-NER はこの固有表現抽出技術を使用して、適切なカテゴリ設計に基づいて作成された初期辞書内の単語から、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書上の性質のことを示す。その特徴量に従い単語抽出が実行され、抽出した単語が該当カテゴリに属するか表す推論確率を 0 から 1 連続値で出力する。推論確率が 1 に近い単語ほど、カテゴリに該当する単語である可能性が高い。

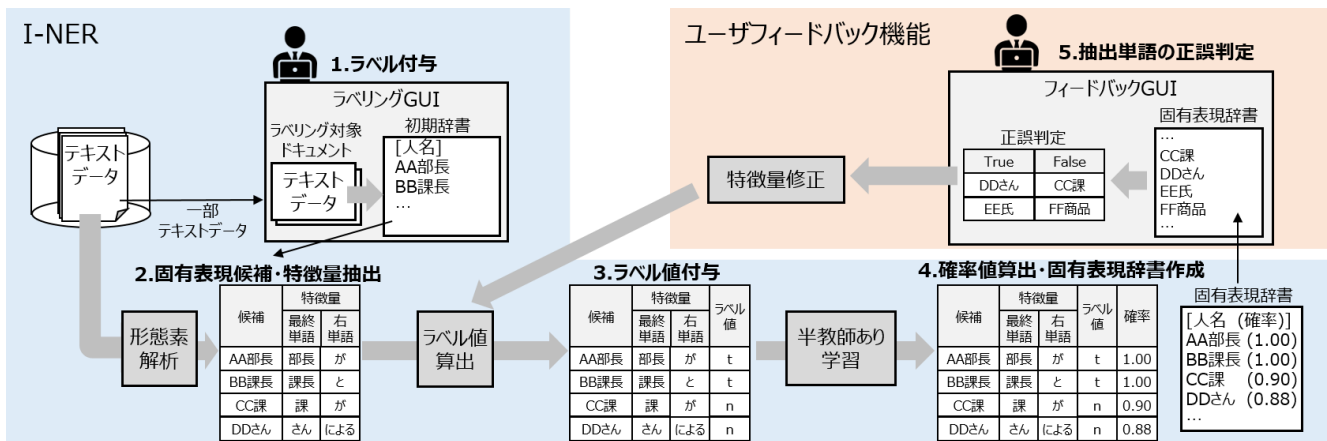


図 1. Smart Dictionary 基盤の実現方式

### 3.2 ユーザフィードバック機能

I-NER にて出力された固有表現辞書の一部の単語をユーザに提示し、ユーザにカテゴリに正しく属しているかの正誤フィードバック情報を入力してもらう。正誤判定結果から特徴量を一部修正し再学習することで、省工数で精度の向上を図る。下記にユーザフィードバック機能の処理概要を記載する。

Smart Dictionary 基盤のフィードバック手順：

Step 1.固有表現抽出を実行 (I-NER)

Step 2. SME がフィードバック作業の終了判断するまで以下を繰り返す

Step 2-1. SME に正誤判定を行ってほしい固有表現の優先順位を算出 (詳細は 3.2.1 節)

Step 2-2. フィードバック GUI を通して SME が正誤判定を実行

Step 2-3. 正誤判定の単語が持つ特徴量の重み値を上げ、誤抽出単語が持つ特徴量の重み値を下げる

Step 3. 固有表現抽出を再実行 (I-NER)

上記手順において、最小限の作業工数で抽出精度改善を達成するためには、Step 2-2 で SME が行う正誤判定数を最小限にする必要がある。そこで少数の正誤判定でも効率的な抽出精度改善を可能にするために、正誤判定する単語を選択する規則を作成した。

#### 3.2.1 正誤判定単語の選択規則

正誤判定数が少ない場合でも効率的に特徴量の修正が可能になるように、以下の 4 つの規則を作成した。SME に正誤判定を要求する優先順位が最も高い固有表現は、[規則 1] もしくは[規則 2]に該当する固有表現である。ただし、両規則ともすでに 1 回以上フィードバックを実行していることが条件であるため、初回フィードバックのみ[規則 3]に基づいて優先順位を決定する。また、[規則 4] は[規則 1] [規則 2] [規則 3]いずれかを適用した場合に、同率順位が発生した際のサブ規則として適用する。

[規則1] 正誤判定済み固有表現の中で保有数が少ない特徴量を多く保有する固有表現を優先して判定

[規則2] フィードバック実行後に確率値が大きく変動した固有表現を優先して判定

[規則3] 確率が 0.5 前後の固有表現を優先して判定

[規則4] 文書データ中に頻出する固有表現を優先して判定

以上 4 つの規則を適宜組み合わせることで優先順位を決定する。最も効果的に特徴量修正が可能で [規則 1] [規則 2] は 2 回目以降のフィードバック時を前提としているため、本提案手法は複数回のフィードバックを行うことで NER 精度改善する手法を前提とする。

## 4. 評価

### 4.1 評価方法

本研究では社内の課題解決事例集文書 169 件から人名を抽出する処理において、ユーザフィードバック機能を使用することによる抽出精度の改善効果を評価した。ユーザが正誤判定する単語については、上記の正誤判定の優先順位の規則に従って選択した場合と、ランダムに単語を選択した場合の 2 通りで評価した。精度評価 (F 値, precision, recall) は、Smart Dictionary 基盤によって抽出した単語 1330

個全ての正誤を予め手動で確認して作成した完全辞書と比較することで求めた。

### 4.2 評価結果

図 2 は、抽出単語 1330 個のうち、20, 40, 60 個の単語のみユーザが正誤判定した場合の抽出精度である。最も良い精度を示したのは、規則に基づいて 60 個の単語の正誤判定を行った場合で、すなわち全抽出単語のうち 4.5% の単語のみに正誤判定を行った結果である。4.5% の単語判定により、フィードバック適用前と比べて F 値が 0.53 から 0.84, precision が 0.45 から 0.90, recall が 0.64 から 0.79 に改善した。

## 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、有用な情報を抽出する際に必要となる構造データ化基盤である Smart Dictionary 基盤におけるユーザフィードバック機能を開発した。本研究では、課題解決事例集からの固有表現抽出に適用し、精度評価を行った。フィードバック適用前は F 値 0.53 に対し、4.5% のみ正誤判定することで F 値を 0.84 に向上することを確認した。今後、他の企業内文書に本基盤を適用することにより、更なる有効性の検証を行う。

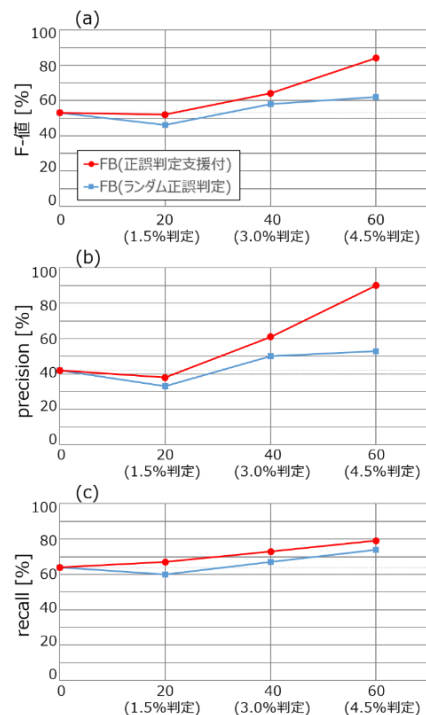


図 2. ユーザフィードバック回数による NER 精度比較 (全抽出単語数における判定した割合)

### 参考文献

- [1] N. M. Kevin and J. K. Ngatia, "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] 照屋絵理, Smart Dictionary 実用化に向けた教師データ量と NER 精度評価について, 第 18 回情報科学技術フォーラム (2019)