E-024

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
Training Data Size and NER Accuracy Estimations for Practical Use of the Smart Dictionary

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を準備するの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]。Smart Dictionary 基盤を用いるにあたり,求める情報抽出精度を得るために distant supervision で必要となる教師データ量が分からないという課題があった。そこで本研究では，Smart Dictionary 基盤を用いて情報抽出を行い,教師データ量と精度の関係性を評価した。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

社会保険報酬支払基金[6]では業務効率化に向け，保険診療審査の自動化を推進している。保険診療審査では，各医療機関の医師が記載する診療報酬明細書（レセプト）と呼ばれる文書を医療機関から受け取り，当該文書に記載された保険診療行為が，自治体が定めたルールに則っているかの審査を実施している。大量のレセプトに記載された様々な保険診療行為の審査には，人手やルールに関する専門知識がいるため，自動化を推進している。しかし，人手で記載されるレセプトには“インスリン”と“インシュリン”など，病名や薬名の表記ゆらぎが大量に存在する。そのため，これらの表現のゆらぎに対応できず，審査の自動化が困難であるという問題が発生した。もし，レセプトから病名や薬名を抽出し，さらに抽出した病名や薬名の関連語/類義語を認識し，表記揺れの辞書を作成することが出来れば，本問題に対応することが可能である。

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。Smart Dictionary 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（非抽出対象）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「AA 症候群を発症した。」という文章内の「AA 症候群」という単語の場合，最終単語という特徴量に対し「症候群」，右単語という特徴量に対し「が発症」

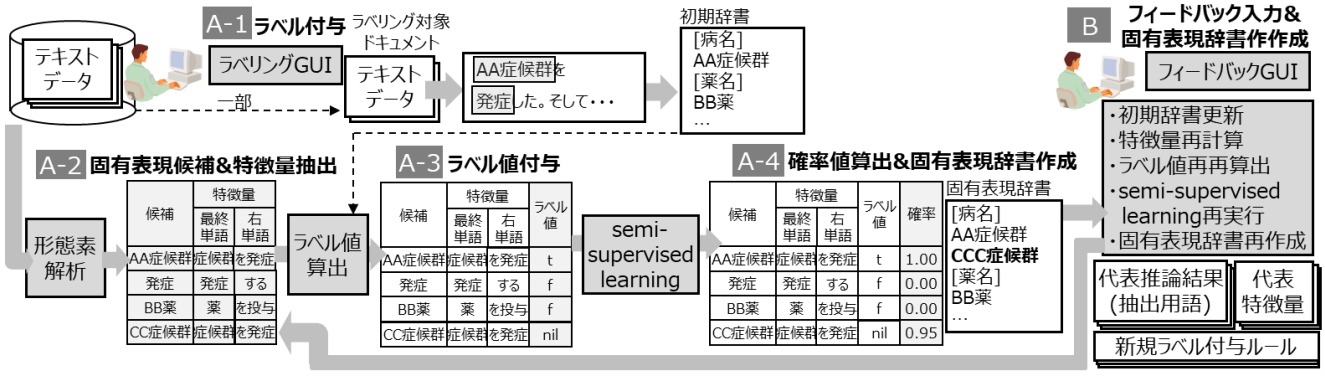

図 1 Smart Dictionary の実現方式

(株) 日立製作所　Hitachi, Ltd.

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 の正誤フィードバック情報を入力してもらい，その結果を初期辞書に加え再学習する。これにより，低コストで精度の向上を図る。

4. 評価

4.1 評価方法

本研究では Medical Information Mart for Intensive Care [7] から取得した退院レポートから病名および薬名を抽出する精度を評価した。評価方法の詳細を下記に示す。

4.1.1 退院レポート数と抽出精度の評価

学習に用いる退院レポート数を変化させた際の精度を評価した。退院レポート内に存在する 169 件の病名および 198 件の薬名を初期正解辞書として使用した。また，退院レポート内に含まれる病名 50 件，薬名 50 件を正解データとし，再現率を評価した。

4.1.2 初期正解辞書内の単語数と抽出精度の評価

学習に用いる初期正解辞書内の単語数を変化させた際の精度を評価した。退院レポート 250 件を学習に用いた。また，退院レポート内に含まれる病名 50 件，薬名 50 件を正解データとし，再現率を評価した。

4.2 評価結果

4.2.1 退院レポート数と抽出精度の評価結果

図 2 は学習に用いる退院レポート数に対する再現率の変化を示した図である。病名・薬名共に再現率が向上し，100 件の退院レポートを用いた場合再現率は 60%程度であるが,500 件の退院レポートを用いた場合 90%に向上した。

4.2.2 初期正解辞書内の単語数と抽出精度の評価結果

図 3 は学習に用いる初期辞書内の単語数に対する再現率の変化を示したグラフである。病名・薬名共に単語数を増やすと精度が向上したが，ある一定（80%程度）で再現率が収束するという結果になった。

4.2.3 結果の比較

退院レポート数および初期正解辞書内の単語数と抽出精度の評価を比較すると，1 件データを増やした際の再現率の上昇率は初期正解辞書内の単語数を増やした際がより大きくなった。このことから，退院レポートおよび初期正解辞書内の単語数を増やす工数が同じだと仮定すると，初期正解辞書内の単語数を増やした方が効率的に再現率の向上が可能であることが分かった。また，退院レポートを増加させた場合は再現率が 90%に達したのに対し，初期正解辞書内の単語数を増加させた場合は再現率が 80%で収束した。このことから，退院レポートを増加させた場合の方が再現率の高い上昇幅が見込まれることが分かった。

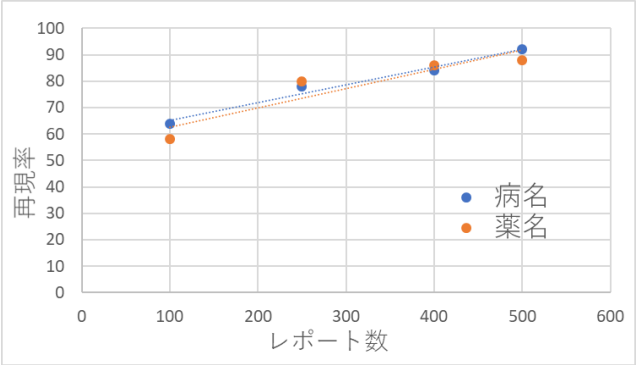

図 2　退院レポート数に対する再現率の変化

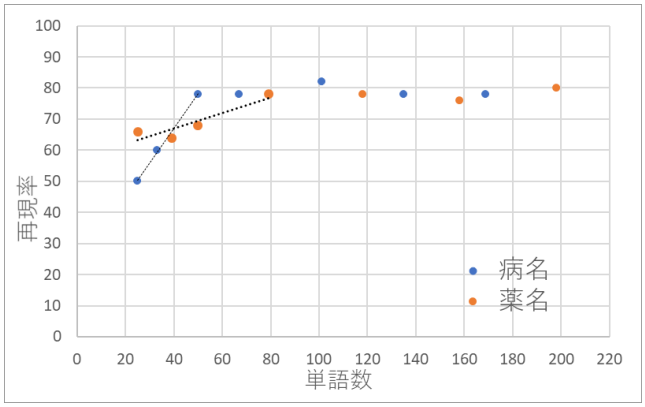

図 3 初期正解辞書内の単語数に対する再現率の変化

5. おわりに

本研究では，テキストデータを構造化し有用な情報（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。本基盤を用いて退院レポートから病名および薬名を抽出する精度を評価した。教師データとして使用する退院レポート数および初期辞書内の単語数と抽出精度の変化の評価を行ったところ，退院レポートおよび初期正解辞書内の単語数を増やす工数が同じだと仮定すると，初期正解辞書内の単語数を増やした方が効率的に再現率の向上が可能であることが分かった。また，退院レポートを増加させた場合の方が再現率の高い上昇幅が見込まれることが分かった。今後，他分野の文書に本基盤を適用することにより，本基盤の有効性および教師データ量と精度の関係性を検証する。

参考文献

[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 (2013)
[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.
[6] 社会保険報酬支払基金, https://www.ssk.or.jp/
[7] MIMIC-III, a freely accessible critical care database. Johnson AEW Pollard TJ, Shen L, Lehman L, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, and Mark RG. Scientific Data (2016).