E-043

弱教師学習による営業日報分析を用いたカタログ検索
Catalog Search using Sales Report Analysis with Distant Supervision

加藤 大羽† 田中 美智子†
Daiba Kato Michiko Tanaka

1. はじめに

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近年注目されてきている[1,2]。このようなナレッジマネジメントの重要性が認識されるのに伴い，営業日報，企画書や提案書，アンケートなど大量の文書データを自動的に分析し，その内容を迅速に把握して営業活動に生かしたいというニーズが高まっている[3]。

しかしながら，企業の業務を通じて収集・蓄積された非構造のテキストデータは，例えば選択形式で収集された構造化済みテキストデータと比べ，データマイニングにおける課題が多い。テキスト系の非構造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，その作業に膨大な時間を必要とする。

また，報告書などは業界用語や社内用語といった一般的でない用語を含んで記載されていることが多いため，AI を用いたテキスト分析の前に，データサイエンティスト（DS）とドメイン知識を持つユーザ（SME）の間で，ドメイン知識をすり合わせる導入コンサルの作業が必要となる。

これらの作業軽減化に向け，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 を開発した。Smart Dictionary 基盤は，ドメイン知識を持つユーザ（SME）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，省工数での単語抽出を可能にし，迅速な非構造データ分析を実現する。

社内用語を多く含む非構造のテキストデータである営業日報を自動的に分析し，最適な自社製品・コンテンツを提案することで，営業活動の情報収集効率化を目指している。営業日報と製品カタログは，記載表現が大きく異なるため，営業日報に記載されている用語をそのまま入力して検索しても，関連する製品が十分にレコメンドされないことが課題であった。

本研究では，Smart Dictionary を用いて営業日報と製品カタログの 2 種類の文書から得たドメイン知識の単語辞書を活用することで，営業日報の内容を自動分析し製品カタログからレコメンドする検索システムを開発した。類義語・関連語抽出技術を活用して，営業日報中の用語をカタログ中によく記載されている表現に変換もしくは拡張することで，本来別々の用途で記載された営業日報と製品カタログの関連性を求めることができる。単語辞書の抽出精度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 基盤の概要と非構造データ分析，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，既存技術では社内用語など独自の単語を含む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。さらに，営業日報と製品カタログの情報連携など，異種文書間では表現方法やよく使われる言い回し等がことなるため，2 種文書間を繋ぐ類義語・関連語辞書が必要になる。省工数でこれらの準備が可能になれば，社内に眠る非構造データの二次利用・連携が容易になり，類似事例等の検索やレコメンドなどへの活用が可能になる。

3. Smart Dictionary 基盤

社内に眠る非構造データの省工数での構造データ化を実現するため，我々は Smart Dictionary 基盤を開発した。Smart Dictionary 基盤は，ユーザが入力する限られた量のドメイン知識情報をもとに低コストかつ高精度な単語抽出を実現する I-NER（Interactive Named Entity Recognzer）と，I-NER で生成された情報を利用することで精度よく類義語・関連語関係の抽出を実現する L-HRE（Label-based Hidden Relation Extractor）を備える。図 1 は，Smart Dictionary による類義語・関連語辞書の生成の概略図である。I-NER は半教師あり学習（semi-supervised learning，もしくは Distant Supervision）により，教師あり学習と比べ低コストで単語抽出を可能にする技術である。L-HRE は，I-NER で単語抽出し生成した単語辞書の中から，教師なし学習を活用して単語間の関連（類義語や関連語関係となる単語のペア）を抽出し，類義語・関連語辞書を生成する。

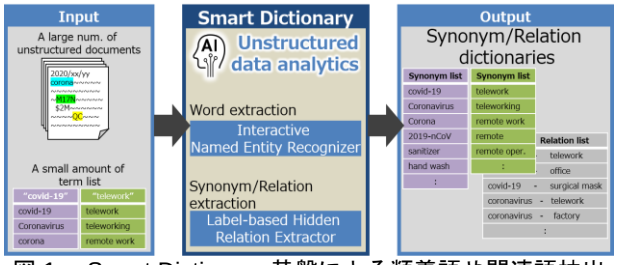

図 1　Smart Dictionary 基盤による類義語や関連語抽出

3.1 重要語抽出とフィルタリング

大量の営業日報を構造データ化し二次利用するためには，製品名や目的などの重要語と，金額や人名などの二次利用すべきでない機密やプライバシーを含む除外単語を認識する必要がある。Smart Dictionary の I-NER による単語抽出は，あらかじめユーザにより定義された任意のカテゴリ（人名・組織名・地名・日付・数量など）へ，文中の単語を分類することで文書構造を理解する。このカテゴリ設計を利用して重要語と除外単語の認識を行う。

(株) 日立製作所　Hitachi, Ltd.

営業日報から単語抽出を行う際，営業活動の目的や顧客が対応したい課題，提案候補の製品などを"重要語のカテゴリ"として設計する。同時に，予算規模や人名，部署名などを"除外する単語のカテゴリ"として設計する。これらのカテゴリ設計に基づき営業日報から単語抽出を行い，"重要語のカテゴリ"で抽出された単語のみを，営業日報から抽出した重要語として製品カタログ検索に利用する。"除外する単語のカテゴリ"を設計していることで，"重要語のカテゴリ"内に予算規模や人名等が誤抽出されることを抑制できる。

3.2 類義語・関連語抽出

Smart Dictionary の L-HRE による類義語・関連語抽出は，I-NER で抽出した単語の中から類義語・関連語の関係を抽出する。これは以下のような利点がある。

- 不要な類義語・関連語の関係を抽出しない
  - I-NER で設計したカテゴリにどれにも属さない，もしくは"除外する単語のカテゴリ"に該当する単語の類義語・関連語を含まない類義語・関連語辞書を作成することができ，辞書整備がしやすい。
- 全ての単語にカテゴリ属性が付与されている
  - これにより，同じカテゴリの単語の関係抽出であれば類義語関係，別々のカテゴリの単語の関係抽出なら関連語関係と，明確に区別が可能になる。
  - 特定のカテゴリ間の関連語のみ表示するといった切り替えが可能になる。例えば[製品名]と[適用業界]の関係の単語ペアのみを確認するなどができる。

本研究では，大量の製品カタログを対象データとし，類義語・関連語辞書を作成した。これにより，本辞書に登録されている単語は検索対象となる製品カタログに含まれる単語となる。

3.3 営業日報の自動分析からのカタログ検索

Smart Dictionary で作成した重要語・除外語辞書と、類義語・関連語辞書を用いて，営業日報から重要語を判定し，カタログ検索に有用な類義語・関連語をユーザーに提案する。図 2 は，カタログ検索までの概要図であり，大きく分けて 3Step で検索キーワードをユーザーに提案する。

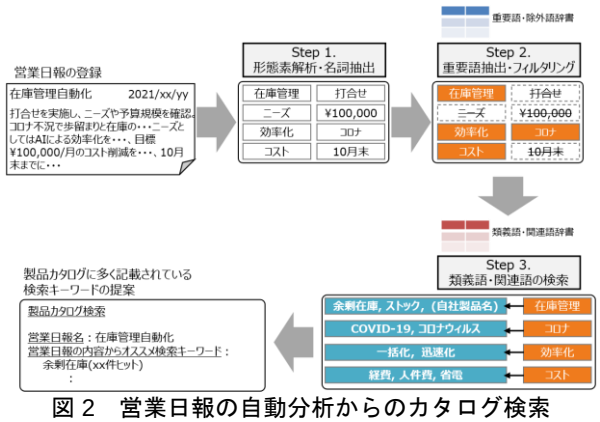

図 2　営業日報の自動分析からのカタログ検索

Step 1. 営業日報の文章を形態素解析し，名詞抽出を行う

Step 2. 抽出した名詞の中から、3.1 節で作成した重要語・除外と辞書と照合し，重要語のみ選択する

Step 3. 選択した重要語について，3.2 節で作成した類義語・関連語辞書でと照合し，検索キーワード候補とする。

検索キーワード候補を製品カタログ中検索でのヒット件数などとともにユーザーに提示することで，ユーザーは営業日報のアップロート作業のみで，関係する製品を検索するキーワードを取得することができる。

4. 評価

本研究では社内営業担当者が実際に作成した営業日報 1876 文書と製品カタログ 650 文書を用いて評価を行った。ユーザーが単語辞書に含まれる単語の一部を正誤判定することで，その情報から抽出モデルを修正するユーザーフィードバック機能を開発し精度改善に用いた[5]。

生成した単語辞書から 60 単語を正誤判定し，単語抽出モデルの修正を行った。単語辞書の抽出精度は F 値 0.53 から 0.84 に改善した。また，この精度改善作業含め辞書モデル生成から異種文書連携までに必要な時間を見積もった。図 3 が結果で，手作業で準備した際 149 時間必要な総作業時間を，本提案手法でおよそ 51 時間（0.3 人/月）で実現する。この結果，0.3 人/月の作業時間で，異種文書間連携による検索システムを実現する見込みを得た。

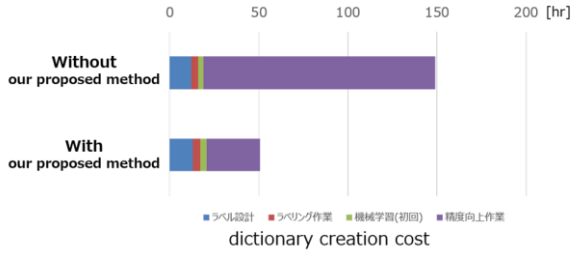

図 3　カタログ検索までの辞書生成の作業工数

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 基盤を用いて，営業日報を自動分析し，オススメ製品の検索支援する異種文書間の連携技術を開発した。重要単語を抽出する際に，自動抽出したい重要単語の指定と，金額や人名等の機密情報やプライバシー情報の恐れがある単語を同時に定義することで，誤抽出を抑えることができることを確認した。その結果，構造データ化作業を省工数化実現の見込みを得た。

参考文献

[1] Antons, David and Grünwald, Eduard and Cichy, Patrick and Salge, Torsten Oliver, "The application of text mining methods in innovation research: current state, evolution patterns, and development priorities", R&D Management, Volume 50 (2020)

[2] Gregory Gimpel, "Bringing dark data into the light: Illuminating existing IoT data lost within your organization", Business Horizons, Volume 63, Issue 4 (2020),

[3] Stefan Wengler, Gabriele Hildmann, Ulrich Vossebein, "Digital transformation in sales as an evolving process", Journal of Business & Industrial Marketing (2021)

[4] 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)

[5] 加藤大羽, ユーザフィードバックによる固有表現抽出の精度改善, 第 19 回情報科学技術フォーラム (2020)