E-048

事業内容文書からの取扱品目に関する情報の抽出
Product Information Extraction from Business Description Documents

川端 篤[1)]　南條 浩輝[2)]
Atsushi Kawabata　Hiroaki Nanjo

1　はじめに

日本標準産業分類は，経済活動を分析するための基盤として広く利用されている．しかし，産業構造の急速な変化に伴い，数年から 10 年単位で改定される産業分類だけでは最新の産業動向を正確に捉えることが困難となっている．このような背景から，更新頻度が高い企業 web ページや SNS から企業の分類を行い，リアルタイムに産業動向をとらえることが望まれている．

本研究では，企業の取り扱う商品という面に着目し，web ページなどの事業内容文書から，テキスト解析 AI により自動的に企業の取扱品目に関する情報の抽出を行い，その有効性を検証する．

具体的には，固有表現抽出を行って取扱品目名を取り出し，次に取り出した取扱品目名のゆらぎを解消し，整理するためにエンティティリンキングを行ったのでその結果について報告を行う．

2　取扱品目の抽出

本研究ではまず，事業内容文書から取扱品目を抽出する固有表現抽出モデルの構築を行う．固有表現抽出モデルには，BERT と生成 AI の両方を用いて精度を比較し，両者の特徴を明らかにする．

2.1　固有表現抽出データセット

固有表現抽出モデルの学習およびテストデータについては，取扱品目に関する情報抽出のための，人手でアノテーションした独自のデータセットを作成する．対象となる事業内容文書のテキストは，株式会社帝国データバンクの保有する信用調査報告書のデータ（991 文書，取扱品目 5,109 件）を用いる．

2.2　BERT

BERT に CRF 層を追加した BERT-CRF [1] を用いる．基盤モデルの BERT は，幅広い大量テキストで学習されている．そこで，本研究の対象となる事業内容に関する文章に特化させるため，事前に追加学習も行う．BERT 基盤モデル[1)]を固有表現抽出データセットでファインチューニングする方法を BERT-CRF とし，後者の BERT 基盤モデルに追加学習してから固有表現抽出データセットでファインチューニングする方法を BERT_ADD-CRF とする．

BERT_ADD-CRF での事前追加学習としては，まず事業内容文書の全データ（1,785,682 文）を用いて，MLM タスクの学習を行い，さらに，wikipedia を用いた日本語の固有表現抽出データセット [2] を用いて，固有表現抽出の学習を行うこととした．

1）　滋賀大学/帝国データバンク
2）　滋賀大学
1）　https://huggingface.co/tohoku-nlp/bert-large-japanese-v2 を用いた

表 1　固有表現抽出 別ドメインにおけるモデルの精度比較

	同ドメイン $P\text{-}f1$	別ドメイン $P\text{-}f1$	変化率 (%)
BERT-CRF	0.9407	0.8077	-14.1
BERT ADD-CRF	**0.9479**	**0.8436**	**-11.0**
LLM-LIST	0.7971	0.7017	-12.0

2.3　生成 AI

生成 AI では，プロンプトエンジニアリングで取扱品目をテキストから抽出する手法（LLM-LIST）を用いる．生成 AI の基盤モデルとして，meta-llama/Llama-2-70b-chat-hf[2)] プロンプトには，出力の事例を含めて生成 AI に入力する few-shot（事例数は 2）を用いる．固有表現抽出データセットのうち，ラベル A (Activity) は，プロンプトによる出力指示が難しく期待する出力が得られなかったので，LLM-LIST では，ラベル P のみ出力される形となる．

2.4　評価方法

事業内容文書においては，製造業やサービス業など業種が異なる場合，出現する単語に大きな違いがある．学習に含まれていない未知の企業の取扱品目に対する予測精度を評価するため，テストデータを，学習データと同じ業種の事業内容文書が含まれる「同ドメイン」，異なる業種の事業内容文書が含まれる「別ドメイン」の 2 種類用意する．評価指標は P ラベルの予測に対する f1 スコア（$P\text{-}f1$）を用いる．評価指標の算出には，文字列の一致率を計算するスパン部分一致を用いる．

2.5　結果

各モデルによる学習・予測を行った結果を表 1 に示す．同ドメイン・別ドメイン共に，BERT_ADD-CRF が $P\text{-}f1$ において最も高い結果となった．同ドメインの $P\text{-}f1$ は 0.9479，別ドメインの $P\text{-}f1$ は 0.8436 と，学習データにはない業種のデータである別ドメインの精度は低下している．しかし，$P\text{-}f1$ の変化率は-11.0%と最も低く，他モデルに比べ精度の悪化は抑えられていることがわかる．BERT_ADD-CRF は，別ドメイン含めた事業内容文章の全体的な傾向を学習しており，それが精度の向上に寄与したものと考えられる．

3　取扱品目の整理

固有表現抽出にて事業内容文書から抽出した取扱品目は，テキストからそのまま抽出したものであるため，表現のゆらぎなどが存在している．そこで，取扱品目を整理するために，辞書（知識ベース）を用意し，抽出した取扱品目を知識ベースをエンティティリンキングにより結び付ける．

2）　https://huggingface.co/meta-llama/Llama-2-70b-chat-hf

3.1 エンティティリンキングモデル

エンティティリンキングのモデルには，2 つの BERT エンコーダー（Context Encoder と Candidate Encoder）から構成される Bi Encoder Model を用いる．Context Encoder には，テキスト中の固有表現と文脈テキストを，文区切りを表す [SEP] でつないで入力する．Candidate Encoder には，知識ベースのタイトルと説明文を，同じく [SEP] でつないで入力する．2 つのテキストは，エンコーダーを通してベクトル化される．固有表現と知識ベースのタイトルが同じものを指す場合に，出力されたベクトル同士の類似度が近くなるように，そうでないときは遠くなるように学習する．

3.2 エンティティリンキングデータセット

エンティティリンキングのモデル学習のためのデータセットには，wikipedia のデータを用いる．wikipedia のページから，各ページのテキスト，テキスト内のリンクが設定されている単語（メンション）の位置，およびリンク先ページのテキストを取得し，エンティティリンキングの学習用データセットとする．

3.3 取扱品目と知識ベースの結び付け

wikipedia のエンティティリンキングモデルを用いて，本研究の対象である取扱品目への知識ベースの付与を行う．知識ベースとして，自動車業界に特化したマークラインズ社の公開する自動車部品一覧 [3] を使用する．自動車部品一覧のデータには，自動車部品の名称およびその説明文が収録されている．事業内容文書は，自動車関連の産業分類に属する企業の文書を用意した．

具体的には，先ほど学習した固有表現抽出モデルで抽出した取扱品目とその説明文のペアを Context Encoder に入力して埋め込み表現（Emb_0）を得る．次に知識ベースに含まれる各自動車部品名と説明文のペアを Candidate Encoder に入力して埋め込み表現（$Emb_i (i > 0)$）を得る．Emb_0 と類似度が高い Emb_i を見つけ，取扱品目を自動車部品名（知識ベース）に紐づける．

3.4 生成 AI によるノイズ除去

固有表現抽出段階では知識ベースの情報を与えていないため，抽出された取扱品目には知識ベースに含まれないものが存在している．これらがエンティティリンキングモデルで知識ベースのエンティティに誤って紐づけられることがある．この誤って紐づけられた取扱品目をノイズとよぶ．

本研究では生成 AI を用いてノイズ除去を行う．具体的には，取扱品目および文脈テキスト，および知識ベースの候補文をプロンプトで与えて，意味が一致するかどうかを判断させることで行う．意味が一致すると判断された場合は，入力した取扱品目と候補文をエンティティリンキング対象として，そのまま保持する．付与した候補文（複数候補文を付与した場合は全て）で，意味が一致しないと判断された場合は，その取扱品目をノイズとして扱い，最終的な取扱品目の出力からは除く．

3.5 エンティティリンキングモデルの評価

評価用データは自動車関連企業の事業内容文書から，固有表現抽出モデルで抽出された取扱品目の 1,000 件とする．この 1,000 件に対して，wikipedia で学習したエンティティリンキングモデルを用いて，知識ベースのエンティティ及びその説明文と紐づける．その際，上位 1 件を紐づける．この 1,000 件のペアに対して，ペアの意味が一致するかどうかを評価した．133 件の意味が一致，867 件の意味が不一致であった．この 867 件はノイズであり，次に生成 AI によってこのノイズ除去を行う．

表 2 エンティティリンキング結果（ノイズ除去後）

	Precision	Recall	f1
Qwen1.5-72B-Chat	0.609	0.842	0.707
ao-karasu-72B	0.615	0.842	0.711
gemma-2-27b-it	0.545	0.820	0.655
calm3-22b-chat	**0.951**	0.579	**0.720**
Qwen2.5-32B-Instruct	0.756	0.466	0.577
Qwen2.5-72B-Instruct	0.551	**0.857**	0.671

3.6 ノイズ除去の評価

生成 AI を用いてノイズ除去を行った．133 件の正しいエンティティリンキング結果をどれだけ見つけられたかの結果を 2 に示す．

calm3-22b-chat が，f1 が最も高い値となった．また，Presicion についても 0.951 と最も高く，ほぼノイズが除かれていることがわかる．実務の観点においては，取引品目とは関係がないノイズを出来るだけ除くことが重要である．calm3-22b-chat は，スクラッチ開発された日本語 LLM であるため，日本語の意味の一致をより厳密に判断していると思われる．ao-karasu-72B は，Qwen1.5-72B をベースに日本語データセットでファインチューニングしたモデルであるが，ベースの Qwen1.5-72B-Chat とほぼ精度は変わらなかった．Qwen2.5-32B/72B は，今回比較したモデルでは最も新しいモデルであるが，本実験においては精度の改善は見られなかった．

4 まとめ

本研究では，テキスト解析 AI を用いて企業の事業内容文書を解析し，固有表現抽出とエンティティリンキングを用いて企業毎に取扱品目を付与する方法を研究した．取扱品目抽出のための固有表現抽出モデルでは追加学習した BERT が適していることを示し，次に抽出した取扱品目候補を BERT と生成 AI により知識ベースにエンティティリンキングする方法を提案した．事業内容文書と任意の知識ベース（取扱品目の辞書）を用意するだけで，自動的に各企業に取扱品目を付与することが可能となり，人手による作業を大幅に減らすことが可能となった．

参考文献

[1] Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. Portuguese named entity recognition using bert-crf. *arXiv preprint arXiv:1909.10649*, 2019.

[2] ストックマーク株式会社. Wikipedia を用いた日本語の固有表現抽出データセット. https://github.com/stockmarkteam/ner-wikipedia-dataset(2025-05 閲覧).

[3] マークラインズ株式会社.『部品辞典』1000 部品網羅！クルマの材料・加工法. https://dictionary.marklines.com/ja/(2025-05 閲覧).