ニュース記事からの企業キーワード抽出

奥田 裕樹　髙橋 寛治
Sansan 株式会社 DSOC
{okuda, ka.takahashi}@sansan.com

1　はじめに

企業が販売する商品や運営するメディア、経営する店舗などの名前は、顧客やユーザが企業のサービスを認知する上で重要である。企業活動においては、企業名を大々的に出すことで一般の認知を得る企業もあれば、企業名よりも特定のサービス名を売り出す「デ・ブランディング」と呼ばれるマーケティング手法で、自社サービスの認知を獲得する企業も存在する。飲料の商品名やオンラインのウェブサイト名、高級ファッションのブランド名、居酒屋の店舗名などはすぐに思い出せても、そこから運営する企業名を想起することは難しいことが多い。

Sansan が提供する、クラウド名刺管理サービス「Sansan」では、社員が交換した名刺を取り込むことで自動でデータ化し、ウェブ上で管理、活用することができる。データベースとして蓄積され続ける名刺情報を活用するには、ユーザの思考に合わせた多様な検索機能が不可欠である。Sansan では、企業名や氏名、部署・役職名、住所などの名刺に記載されている情報で検索が可能なほか、企業に関連するキーワードをデータベース内部で保持することにより、名刺に書かれている内容以上の情報で検索することができる。このようなシステムを構築するためには、あらゆる企業のあらゆるキーワードを収集し蓄積し続けなければならない。しかし、企業活動により日々新しく作り出されるキーワードを人手で網羅的に収集するには、莫大なコストがかかる。

そこで、インターネットで配信される文書の中から、企業活動に関連するキーワードを自動で抽出することを考える。ユーザへの周知のために企業は自社ホームページやプレスリリースを通じて企業活動を宣伝し、ニュースメディアも様々な形で経済に関係する企業の動向を配信する。そうした文書から自動でキーワードを抽出することができれば、企業を検索するためのシステムを低コストで維持することができ、長期的な価値提供が可能になる。

本稿では、ニュースメディアが配信するニュース記事から企業のキーワードを抽出する方法を提案する。ニュース記事やプレスリリース中に登場する企業キーワードは鉤括弧で囲まれることが多いことに着目し、ルールにより自動で抽出した企業キーワード候補に対して、意図した企業キーワードかどうかの二値分類の問題として扱い、アノテーションにより作成した正解データに対して精度評価を行った。

2　企業キーワードの抽出

2.1　企業キーワードの定義

まず本稿における「企業キーワード」を定義する。企業キーワードとは

企業活動の中で生まれたモノやサービスを表す名称

とする。例えば、飲料メーカーが自社商品の飲料に付けた商品名はキーワードとして適切であるが、「スポーツドリンク」のような一般的な用語は特定企業とは紐付かないためキーワードとしては不適切である。収集すべきキーワードを分類した結果、表 1 のカテゴリに合致する名称を収集対象とした。

2.2　対象とするニュース記事

抽出対象のニュース記事は、Sansan 株式会社が提供する「Sansan」[1]および「Eight」[2]においてユーザに配信している文書を用いた。取得元のニュースメディアには、主要新聞社や通信社のほか、プレスリリース配信サイト、企業が自社ホームページで配信するプレスリリースなどが含まれる。また Sansan のニュース配信ロジックとして、ユーザが過去に名刺交換をしたこと

[1] https://jp.sansan.com/
[2] https://8card.net/

表 1: 企業キーワードの分類

カテゴリ	説明
サービス・システム・アプリ名	オンライン上に存在するサービスやアプリケーション、データベースやソフトウェア、ツール等の名称
物理的に存在する商品・製品名	食品や日用品、電化製品など、物理的に存在する商品名
Web サイト・オンラインショップ名	Web サイトやオンライン上にあるショップ名
ブランド名	食品やファッション、車などに利用されるブランド名
運営する施設名	アミューズメントパーク、レストラン、マンション、ビル、ショップ などその企業が運営する施設名
拠点名	その企業の社員が営業・活動している窓口や拠点名
イベント・セミナー・キャンペーン・事業名	イベント・講座（セミナー）・キャンペーン・事業やプロジェクト名
テレビ番組・雑誌名	テレビの番組名や、雑誌、カタログ、Web マガジン名
企業名（関連会社）	親会社や子会社、関連のある企業名

のある企業のニュースを配信するため、企業名が少なくとも 1 社以上含まれている記事のみを対象とした。

以下に実際のプレスリリースの記事[3]を示す。

タイトル 名刺アプリ Eight、企業の課題解決を後押しするビジネスイベント「Meets」を発表 〜ビジネスの「買いたい」と「売りたい」をつなぐ〜

本文 Sansan 株式会社は、同社が提供する名刺アプリ「Eight」から、ビジネスイベント「Meets（ミーツ）」が提供されたことを発表します。Meets は、Eight のテクノロジーを活用し、サービスを「買いたい人」と「売りたい人」とをつなぎ、社会の生産性を上げるビジネスイベントです。

ここで企業キーワードとして抽出が期待されるのは「Eight」および「Meets（ミーツ）」である。

2.3 企業キーワード候補の抽出と判定

まず、ニュース記事から企業キーワード候補を列挙する。ニュース記事のタイトルと本文を結合した文書を対象に、鉤括弧（「」および『』）で囲まれた文字列を抽出する。記事の中には文の途中で途切れているものが存在するため、鉤括弧が正しく閉じられていない文字列は対象としない。

次に、列挙された企業キーワード候補が、抽出目的としている企業キーワードであるかの判定を行う。一般に鉤括弧は、商品名などの固有名詞に使用されるほか、会話文や強調表示など様々な用例が存在する。抽出された候補文字列が、企業キーワードかそれ以外かの二値分類を行う機械学習モデルを構築する。

鉤括弧内の文字列に対して企業キーワード名として適するかどうかの二値分類のタスクではあるが、その判定を行うには前後の文脈を考慮する必要がある。人名や時間、数量といった表記がある程度定まった情報に対しては、辞書の整備やルールを定義することにより判定可能ではあるが、企業キーワード名は多様な表現が存在し、名称単体でそれが企業キーワード名と判定するのは人間でも不可能である。そのため、系列モデルとして文脈を考慮することができる固有表現抽出の手法を用いた。

3 実験および評価

3.1 実験設定

2019 年に Sansan にて配信されたニュース記事の中から、全 3,978 件の記事に対してアノテーションを行った。鉤括弧を抽出することにより得た企業キーワード候補 7,225 件のうち、企業キーワードに適するものは 4,439 件、適さないものは 2,786 件だった。学習データ、開発データ、テストデータの比率が 8:1:1 になるように、企業キーワード候補単位でデータを分割した。データセット内には企業キーワードの重複が存在する

[3] https://jp.corp-sansan.com/news/2019/meets.html

表 2: IOB2 タグの付与例 (O タグは省略)

名刺アプリ Eight、企業の課題解決を後押しするビジネスイベント「Meets <B-KEYWORD-1> 」を発表〜ビジネスの「買い <B-KEYWORD-0>たい<I-KEYWORD-0>」と「売り<B-KEYWORD-0> たい<I-KEYWORD-0>」をつなぐ〜

が、件数が少なかったため分割時に各データセット内での異なり数は考慮していない。

また、固有表現抽出のためのデータ整形を行った。アノテーションされた文書に対してキーワード名として適する単語列には「{B,I}-KEYWORD-1」を、適さない単語列には「{B,I}-KEYWORD-0」という IOB2 タグを付与し、鉤括弧を含むその他の単語には「O」タグを付与した。付与例を表 2 に示す。

3.2 モデル

固有表現抽出で用いられる BiLSTM-CRF のモデルにおいて、前後の文字単位の文脈を考慮する Contextual String Embeddings(CSE) による系列モデル [1] を使用した。これは分かち書きした単語列を入力とし、対象単語までの前後の文脈を文字単位で BiLSTM の系列として扱い、単語単位で潜在表現を得ることで各単語に対して IOB2 タグの推定ラベルを得るモデルである。

ベースラインとしてすべての鉤括弧を企業キーワードとする majority class を採用した。また、固有表現抽出との比較として、対象となる企業キーワード候補の前後 10 単語の Bag-of-Words 表現を用いて SVM による二値分類を行った。モデルの学習時には、学習データ内で 5 分割交差検定によるパラメータチューニングを行った。

3.3 評価方法

モデルの評価には、抽出した企業キーワード単位での二値分類における適合率、再現率、F 値を評価尺度として用いた。固有表現抽出モデルでは各単語におけるタグが出力されるため、複数の単語系列のタグから一つのラベルに変換する際には、候補単語列のうち付与数の多いラベルを候補キーワードにおけるラベルとした。

表 3: 企業キーワードの二値分類の評価

Method	Precision	Recall	F_1
majority class	0.31	0.50	0.38
BoW+SVM	0.75	0.72	0.73
BiLSTM-CRF+CSE	0.87	0.82	0.83

3.4 結果

企業キーワード候補の二値分類に対してモデルを適用した結果を表 3 に示す。

それぞれの評価尺度において、BiLSTM-CRF+CSE が最も良い性能を示した。前後の単語の BoW 表現を用いた BoW+SVM と比較して良い評価となったことから、前後に出てくる単語に加えて広く文脈情報も必要であることがわかる。

3.5 エラー分析

BiLSTM-CRF+CSE による判定において企業サービス名であると誤判定した記事の中で、該当の鉤括弧部分の抜粋を示す。

- “新 iPhone「**タッチペン**」対応の可能性、専用ケース準備中の業者”[4]
- “「**新聞×ＡＲ**」の表現アイデアコンテスト”[5]
- “ワークフローを電子化する「**ワークフローシステム**」を展開している”[6]

これらは、強調する意図で用いられている事例である。単語の意味から明らかに異なるとわかるものもあれば、文脈だけでは一概に判定できない場合も存在する。

次に、企業キーワード名ではないと誤判定した記事を示す。

[4] https://forbesjapan.com/articles/detail/29451
[5] https://prtimes.jp/main/html/rd/p/000000198.000011115.html
[6] https://m.finance.yahoo.co.jp/news/detail/20191001-00000004-scnf-stocks

- “日刊工業新聞社発行の月刊誌「**工場管理**」１１月号では”[7]
- “「**車いすで仲間と一歩外へ**」を始動、サッカー観戦における車いす席の稼働率を高める取り組みを実施”[8]

こちらは逆に一般名詞ではあるが雑誌名として成り立っているものや、フレーズのような形式ではあるがキャンペーン名として用いられている事例であった。人間であれば前後の文脈から総合的に判断可能である一方で、ニュース記事のタイトルには冒頭に鉤括弧が来る場合も多く、利用できる文脈情報に制限がある場合もあった。

4 関連研究

この章では企業キーワード抽出に関連する研究について述べる。他の言語リソースからの企業キーワード抽出として、ブログ記事を対象にした研究がある。渡邊ら [3] は、ブログ上のレビューから商品名を抽出する手法を提案している。また、池田ら [5] は、ブログ記事から土産の品名および販売店舗名といった特定の表現を対象に固有表現抽出を行っている。これらの研究は構造化されていない文書からから情報抽出するという意味で、本稿の研究と似たモチベーションである。

より網羅的にあらゆる情報を構造化する取り組みとして、「森羅：Wikipedia 構造化プロジェクト」は、Wikipedia の記事文書に対して関根の拡張固有表現を構造化した形で抽出するリソース構築プロジェクトである [2]。拡張固有表現で定義されている名前表現には「1.7 製品名」というグループが存在し、この中で企業活動に関連する項目が企業キーワードに相当する。Wikipedia 内には企業活動に関連する記事が多数存在するため、企業キーワードを取得する上で有用な情報となるが、一方で Wikipedia に作成される項目は企業やサービス名が広く一般に認知されている対象に限定され、新しく誕生した企業キーワードに対しては反映に時間がかかる場合がある。そのため、幅広く最新の情報を取得する本稿のような場合の利用には適さない。

また、オンラインショッピングサイトの商品説明文からの属性値抽出 [4] において、本稿で対象とした企業キーワードが含まれる場合がある。対象とする属性値の中には企業キーワードと扱うことのできる商品名等が含まれているが、一方で商品の説明文から抽出する項目は、人間が想起するようなキーワードよりも品番や正式名称を主に対象にしている。明確に記載がされている文書からの情報抽出という点で、あらゆる記事が配信されるニュース記事とは対象とする文書の性質が異なる。

[7] https://www.nikkan.co.jp/articles/view/00534784
[8] https://prtimes.jp/main/html/rd/p/000003593.000003442.html

5 まとめ

本稿では企業活動に関連するキーワードを自動で抽出する手法を提案し、独自にアノテーションしたニュース記事に対して適用した。今後は、さらにデータを拡充させることで精度向上および多様な文書への適用を可能にさせるほか、継続的に運用することにより企業キーワードの辞書拡充を行っていく予定である。また、実用においては抽出した企業キーワードがどの企業に紐づくかを判定する必要がある。そのため、抽出された企業キーワードと、文書中に現れる企業名との Entity Linking を行う必要があり、この処理の自動化も課題である。

参考文献

[1] Alan Akbik, Duncan Blythe, and Roland Vollgraf. Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pp. 1638–1649, 2018.

[2] 関根聡, 小林暁雄, 安藤まや. Wikipedia 構造化プロジェクト「森羅 2018」. 言語処理学会第 25 回年次大会, 2019.

[3] 渡邊尚吾, 乾孝司, 山本幹雄. カテゴリ情報を利用した blog 記事からの商品名自動抽出. 言語処理学会第 19 回年次大会発表論文集, 2013.

[4] 新里圭司, 関根聡. 商品説明文からの属性・属性値の自動抽出. 言語処理学会第 19 回年次大会発表論文集, 2013.

[5] 池田流弥, 安藤一秋. 深層学習によるブログ記事からの土産の品名・店名抽出. 言語処理学会 第 25 回年次大会 発表論文集, 2019.