指導教員: 湯本高行准教授

キーワードの共起語を用いた関連企業の検索

兵庫県立大学社会情報科学部

社会情報科学科

2019 年度入学

JB19S046

辻田 隆善

2023 年 2 月提出

キーワードの共起語を用いた関連企業の検索

JB19S046 辻田 隆善

指導教員: 湯本高行准教授

概要

現代において、インターネット普及などによるテキストデータの増加により、多様な情報を収集することが容易になっている。そのため、就活や投資などの企業判断においてもテキストデータを用いた情報収集が有効な手法である。しかし、半導体不足やコロナなど、とあるキーワードに影響を受けている企業を調べる場合、有価証券報告書やニュースから情報を収集してから判断する必要があるため、一つ一つ調べるには膨大な時間がかかってしまう。

そこで本論文では、キーワードに少しでも影響を受ける企業を検索できる手法を提案する。具体的な手法の概要としては、クエリベクトルの抽出、企業ベクトルの構築、コサイン類似度による関連企業検索に分ける。キーワードから類似度が高い語句を抽出し、その語句を含む記事を関連記事として抽出する。そこから確信度や IDF を用いて共起語を取得し、キーワードと共起語からクエリベクトルを構築する。また、各企業の事業内容を IDF などの条件から名詞を抽出し、企業の特徴ベクトルを構築する。この二つのベクトルのコサイン類似度を求め、上位の企業を関連企業として表示するのが本手法である。

複数のキーワードでの類似度上位 50 件の適合率を求め、実際に上位にきた企業が関連企業かどうかの評価を行ったところ、「半導体」「木材」「旅行」「音楽」「ゴム」「カメラ」の適合率は、1.00, 0.88, 0.94, 0.92, 0.72, 0.85 であった。事業内容にキーワードの語句を含んでいない企業だけの上位 20 件では、1.00, 0.85, 0.90, 0.85, 0.65, 0.94 だった。また、事業内容文で共起語を求めた場合の関連企業検索手法との比較も行い、ニュース記事から共起語取得を行った方が幅広い事業の関連企業を求めることも確認できた。しかし、少ないマッチでも上位にきてしまったり、一部の語句の適合率が悪いため、共起語の抽出

方法についても考える必要がある。また、企業によって事業内容の記述量の違いによって関連企業でも上位に挙がっていない企業もある。今後の課題としては、このような問題を解決していく。

目次

第 1 章

序論

現代において、インターネットの普及や SNS の利用などによるテキストデータの増加により、多様な情報を収集することが容易になっている。このようなテキストデータを用いた情報収集は、就活や投資などにおける企業判断おいても有効な手法となっている。就活では、例えば、コロナウイルスの影響などにより、業績が悪化した企業があった場合、将来的にその企業の成長性に関する判断材料となる。そのため、有価証券報告書や決算短信から事業内容や業績を確認したり、ニュース記事から情報収集を行う。また、半導体不足のニュースなどの業界動向を踏まえ、製造業を担う企業への就職意向などに影響することも考えられる。また、投資においても同様に考えられる。過去の業績や財務状況を確認したり、ニュース記事で悪い影響を受けた企業を確認するために情報収集を行う。企業の情報はニュース記事や有価証券報告書、決算短信などから取得できる。しかし、コロナウイルスや半導体不足などによって影響を受けた企業を調べる際に、関連する企業を一つ一つ内容を把握するには時間がかかり、正確な判断が困難になってしまう。また、このような状況を対処するために、企業検索サイトでキーワードを検索して関連がある企業を調べる方法も、便利な手法の一つである。しかし、キーワード検索の場合、例えば「半導体」で検索した際に、半導体製造メーカーなどの専門的な企業が多く、半導体に関連するすべての企業を検索することができない。そのため、キーワードに少しでも関係している企業までを検索できるようにすることが必要である。

そこで本論文では、キーワードに少しでも関係している企業を検索できるようにするための関連企業抽出の手法の提案を行う。具体的な手法としては、キーワードの類似度が高い語句の関連記事を抽出し、そこから共起語を求める。その共起語を用いたクエリベクトルと、企業の事業内容から名詞を抽出した特徴ベクトルのコサイン類似度を求め、上位の値を関連企業として判断する。

本論文の構成は以下のとおりである。1 章では、本論文の序論として、本研究を行った背景や目的について記述している。2 章では、基本事項である CoARiJ、単語分散表現、共起分析の説明と、本論文の関連研究についてを記述している。3 章では、本手法の概要についてを記述し、クエリベクトルの構築方法、特徴ベクトルの構築方法、類似度上位の結果表示方法についてを詳細に記述している。4 章では、使用するデータ、実験方法についてを記述し、実際の評価を行った結果と、その考察について記述している。5 章では本論文の結論を記述している。

第 2 章

基本事項と関連研究

2.1 CoARiJ

本手法では事業内容を用いて企業の特徴ベクトルを用いるために、CoARiJ データセット [1] を用いる。CoARiJ は、自然言語処理で企業分析を行うために、有価証券報告書や CSR 報告書、統合報告書の情報や数値情報をまとめた日本語の企業文書データセットである。このデータセットには、事業内容をまとめたデータセットがあるため、今回はそのデータセットを利用する。なお、事業内容は有価証券報告書や決算短信に記載されている。

2.2 単語分散表現

数種類の単語をベクトル化する際に、単語の数種類の次元数を用意したベクトルで単語を表現し、表現したい単語があれば、対応する次元を返すという仕組みを one-hot 表現という。本手法では、クエリベクトルの構築や、企業の特徴ベクトルの構築に使用している。しかし、この表現は単語の意味をベクトル化しているわけではなく、単語に番号付けを行っていることになる。単語や文章を低次元の実数値ベクトルで表現することを分散表現という。分散表現では、単語の意味を数値化し、意味関係を見ることができるため、文章をベクトル化した際に文脈が近い意味を持つ文章同士は、近いベクトルで表現する

ことができる。

分散表現では、単語や文章を少ない次元数で実数値のベクトルで表現する。文章の意味をベクトル化する前段階として、単語の意味をベクトル化することがあり、これを単語埋め込みや単語分散表現と言う。この二つは指し示す概念が少し異なり、単語埋め込みは、単語の意味をニューラルネットワークが用いる実数空間に「埋め込む」と言う状況に焦点を当てているが、単語の分散表現は、単語を複数の要素からなる実数値で表現し、それらの要素は他の単語の表現にも用いるというアイデアを表す用語である。単語を高次元のベクトルで表現することで単語間の意味関係を見ることができる。「旅行」と「ツアー」と「東京」で例を考えた場合、「旅行」と「ツアー」は意味が近いため、近いベクトルになるが、「旅行」と「東京」は遠い意味になるため、ベクトルは離れた結果になる。分散表現には Word2Vec や fastText などのモデルがあるが、本手法では、単語ベクトルから類似度が高い語句を取得するために、fastText で作成した学習済み日本語モデルを使用している。fastText[2] [3] は、Word2Vec を基に作成されており、この学習済み日本語モデルを使用することで、キーワードとの類似度を一つ一つ計算することなく類似度が高い語句を抽出することができる。なお、Word2Vec[4] は、単語をベクトルに変換する単語分散表現を学習する複数の分析モデルの総称である。基礎モデルには、文章中の単語を周辺の単語から予測するモデルの CBOW と一つの単語から前後の単語を予測するモデルの Skip-gram があり、また 3 通りの高速化の合わせた計 6 種類のモデルがある。なお、CBOW に基づく分散表現は、単語の周辺語から単語を予測するネットワークを学習し、埋め込み層の重みを単語のベクトルだと解釈して用いるものである。

2.3 共起分析

共起分析とは、テキストデータ内の単語などが一緒に登場する頻度を分析する方法である。例えば、「AI」という語句を含む文書の中に、「機械学習」という語句はよく使用されるが、「東京」という語句はあまり使用されない。このように単語同士の共起頻度を算出することで、単語間の関連性を取得することができる。この共起分析は自然言語処

理タスクやテキストマイニングタスクにおいて重要な技術の一つであり、テキストデータから有益な情報を抽出する上で活用される。本手法では、キーワードに関連する語句を取得するために、共起分析を行っている。また、共起性を求める方法にはいくつかの方法があるが、今回共起性を表す尺度として、確信度が高いほど共起性が強いと判断する。購買データなどから、相関ルールを見つけることができる分析方法にアソシエーション分析 [5] がある。例として、「AI」を含む記事には「機械学習」が出現しやすいという相関ルールを見つけることがアソシエーション分析である。このように、「A ならば、B である」というルール $(A \Rightarrow B)$ を発見する方法を分析する。この分析の計算式として確信度があり、今回はその相関ルールの考え方を語の共起分析に導入している。確信度は「AI」を記事に含む記事が「機械学習」も記事に含まれている割合のことを表す。確信度は以下の式 2.1 で算出する。なお、「AI」と「機械学習」をそれぞれ A,B としている。

$$\text{確信度 } (A \Rightarrow B) = \frac{A \text{ と } B \text{ を含む記事数}}{A \text{ を含む記事数}} \tag{2.1}$$

本手法では、上式の値が大きいほど、共起の度合いが強いとしている。また、上式では $A \Rightarrow B$ の確信度と $B \Rightarrow A$ の確信度では値が異なるため、今回はその両方の値を求めて、共起語を抽出している。つまり、「AI」を含む記事は「機械学習」も含んでいるかだけでなく、「機械学習」を含む記事は、「AI」も含んでいるかを確認する。

なお、今回の手法では、共起語は文書を一単位としている。たとえば、「半導体」を含む記事に「自動車」という語句が複数含まれている場合と、一度だけしか使われない記事は同等のものとして判断している。

2.4 関連研究

共起語を用いた関連研究として、間瀬らは WWW ページからキーワードの共起語を自動抽出する実験を行っている [6]。ここでは、共起語の共起尺度に確信度を用いている。また、新谷らは単語の共起頻度と出現位置から新聞の関連記事の検索手法を提案している [7]。ここでは、記事中の各単語に使用頻度や IDF などの重み付けを行うことで、記事

を特徴付ける単語の重みを高くなるようにしている。また、関連企業検索手法の研究として、平野らはテキストマイニングに着目した、テーマにおける関連銘柄抽出手法を提案している [8]。しかし、この手法では、少しでも関連がある企業までを検索結果に表示することを目的としていないため、本研究では少しの関連でも検索できるような手法を提案する。

第 3 章

キーワードの共起語を用いた関連企業検索

3.1　手法の概要

本手法ではクエリとして与えられたキーワードに共起語を加えたクエリベクトルと、企業ベクトルの類似度を計算し、類似度が高い企業から検索結果を表示する。また、処理の流れは以下の通りである。

1. キーワードの共起語の抽出とクエリベクトルの構築
2. 企業ベクトルの構築
3. コサイン類似度による関連企業検索

まず、ニュース記事を用いて、入力されたキーワードに関連する語句を抽出し、クエリベクトルの構築を行う。次に、各企業の事業内容の特徴ベクトルを構築し、それらの類似度を計算して、より広義な関連性がある企業を検索できるように目指す。検索までの概略を図 3.1 に示す。キーワードのクエリとして「半導体」が入ってくると、「半導体」の類似度が高い語句集合を取り出し、その語句を含む記事を抽出して関連記事とする。その関連記事から確信度や IDF による共起語選別を行い、キーワードの共起語集合を抽出

し、ベクトル化を行う。また、事業内容から各企業の特徴を表す名詞を、IDF などによる名詞選別を行い抽出して、その名詞集合をベクトル化する。このクエリベクトルと、各企業の特徴ベクトルの cos 類似度を計算することで、上位を検索結果にする。

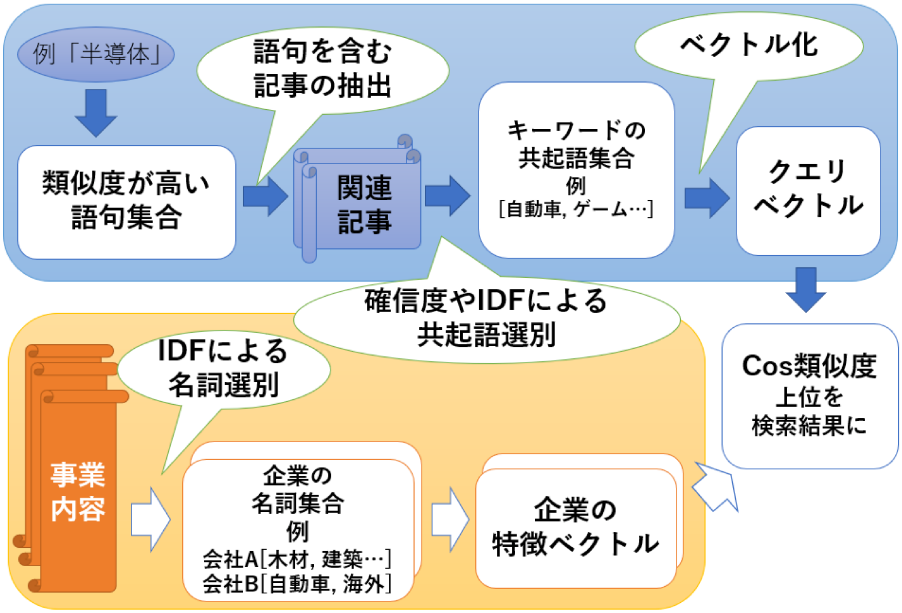

図 3.1　処理手順の概略

3.2　共起語を用いたクエリベクトルの構築

3.2.1　キーワードに関連する記事の抽出

キーワードに関連する記事の抽出を行う。ここでは、キーワードを含む記事と、キーワードと単語ベクトルの類似度が高い値になる語句を含む記事を関連記事としている。なお、ここでの単語ベクトルは fastText で作成した学習済み日本語モデル[*1]を使用しており、キーワードとの類似度が上位の語句を取得することができる。

今回は、キーワードを含んでいない記事も、キーワードとの類似度がしきい値以上の語句が記事に含まれている場合は、関連記事とする。例として、キーワードを「木材」で考

[*1] https://dl.fbaipublicfiles.com/fasttext/vectors-crawl/cc.ja.300.vec.gz

える。キーワードの「木材」という語句を含む記事以外に、「木材」との単語ベクトルの類似度が高い「材木」や「製材」などの語句を含む記事も、関連記事に含まれている。これにより、キーワードを記事に含まないが、キーワードと関連が近い記事も関連記事として抽出することができる。キーワードを「木材」で考えた際の関連記事の抽出方法の流れを図 3.2 に示す。ここでは、同義語の意味を持つような語句を抽出できるようにしているため、しきい値を 0.65 としている。そのため、キーワードによっては 0.65 以上の類似語がない場合もある。

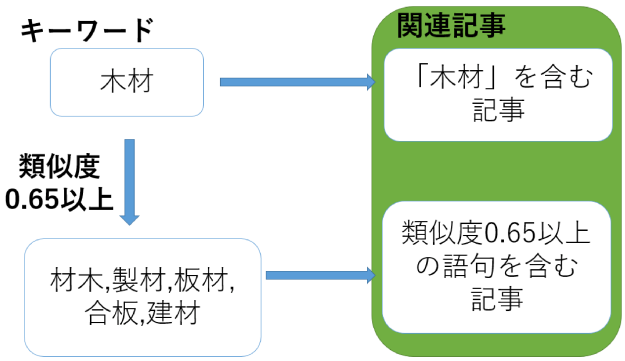

図 3.2　関連記事の抽出方法

3.2.2　関連記事から共起語の取得

共起語を取得するために、特定の語句や企業と関連しにくい名詞は除外し、キーワードと関連する記事から確信度の高い語句を取得する。共起語を選別するために以下の語句を共起語の候補から除外する。

- 名詞と固有名詞以外の語句
- 一文字、人名や地名でない
- 助数詞可能名詞、サ変可能名詞でない
- 記事特有の語句

削除条件の例を表 3.1 に表示する。この条件を満たした各語句の確信度を計算し、上位

の語句 200 件を取得する。IDF などにより多くの語句が削除されたり、キーワードに少しでも関連がある語句を取得できるように多くの値を取得するようにしている。そのため、比較的多い語句を取得している。また、キーワード q と共起語候補の名詞 c としたときの $q \Rightarrow c$ の確信度の定義式を式 3.1 に示す。

表 3.1　名詞の削除条件

削除条件	例
一文字	中, 者
人名や地名	山田, 東京
助数詞可能名詞	人, 枚, 匹
サ変可能名詞	活用, 提供, 拡大
空白を表す語句	￥n, ￥3000, ￥u
記事特有の名詞	株式会社, 以下, 本社, 資料

$$conf(q \Rightarrow c) = \frac{q \text{ と } c \text{ を含む記事数}}{q \text{ を含む記事数}} \tag{3.1}$$

確信度の上位を採用することで、キーワードに一定の共起語候補を採用することができる。確信度をしきい値で採用する方法も検討したが、キーワードごとに関連する記事数が異なるため、キーワードによって共起語の候補の数が問題になる。そのため、今回はこの方法を採用する。

3.2.3　共起語候補の選別

次に、先ほど求めた確信度が上位の語句から、キーワード q に対して、以下の条件を満たす語句 c を共起語とした。

- IDF が 2 以上の語句
- $c \Rightarrow q$ の確信度が上位 40 件の語句

IDF では、どの記事にも使われるような語句を削除する。複数のキーワードから IDF を計算した際に、2 より小さい語句はあまり意味のない語句が多かったため IDF が 2 以上の語を採用する。また、キーワード q と共起語候補の語句 c に対して、$c \Rightarrow q$ の確信度を求める。これは、3.2.2 で述べた $conf(q \Rightarrow c)$ の条件と合わせて、キーワードと共起語の候補が互いに関連していることを表現するためである。しきい値を設けずに上位 40 件にした理由は、しきい値で共起語を抽出すると、キーワードによって関連記事の数が変わるため、キーワードによって確信度の値が極端に変化してしまうためである。また、40 件より少ない場合、キーワードに少しでも関連のある語句が少ない。以上の理由による上位 40 件に採用している。この 2 つを式で表すと以下のようになる。

$$IDF(t, D) = \log \frac{N}{1 + df(t)} \tag{3.2}$$

$$conf(c \Rightarrow q) = \frac{c \text{ と } q \text{ を含む記事数}}{c \text{ を含む記事数}} \tag{3.3}$$

なお、IDF での t は共起語候補の語句、D は記事集合、N は全記事数、$df(t)$ は語 t を事業内容に含む記事数としている。

3.2.4 クエリベクトルの構築

先ほど求めたキーワードの共起語とキーワードを含めた集合を用いてクエリベクトルを作成する。この集合をベクトル化する際に、IDF で重み付けを行う。IDF を用いる理由としては、企業で多く使われている語句が上位に来てしまうのを防ぐためである。また、キーワード等の頻繁に使用される語句を含むと上位にきてしまうため、少しでも関連のある企業が上位にくるように IDF を用いた。また、今回重複する語句は削除しており、TF を用いても全て 1 になるため、ここでは使用していない。

3.3　企業の特徴ベクトルの構築

3.3.1　企業の特徴を表す名詞の抽出

事業内容から、企業ごとに形態素解析を行い、企業の特徴を表す語句を抽出する。ここでの特徴を表す語句は、名詞と固有名詞から、特定の条件を満たした語句を、採用している。以下に条件を記述する。

- 表 3.1 の削除条件に当てはまらない名詞
- IDF が 1.3 以上

IDF がしきい値以上の語を削除する目的は、どの企業にもよく使われるような名詞を削除し、より企業の特徴を表す名詞を抽出するためである。これらの条件を満たした名詞を、企業の特徴を表す名詞として抽出する。IDF を 1.3 以下の値には、企業の特徴を表すような語句が少なく、どの企業にも関連があるような語句が多かったため、今回は 1.3 以上にしている。

3.3.2　企業ごとの特徴ベクトルの構築

企業の特徴を表す名詞のみを含めた集合から、特徴ベクトルを作成する。ここでもクエリベクトルと同様に、IDF を用いて、ベクトルの構築を行う。なお、今回の事業内容でも特徴を表す名詞が重複しないように削除しているため、TF は用いていない。

3.4　類似度上位による結果表示

先ほど求めたベクトル同士のコサイン類似度を求め、類似度が高い企業を関連企業として、類似度の上位何件を検索結果に表示する。クエリベクトルを v_q、企業の特徴ベクトルを v_c とすると、コサイン類似度は式 3.4 で算出される。

$$cos(v_q, v_c) = \frac{v_q \cdot v_c}{|v_q|\,|v_c|} \tag{3.4}$$

第 4 章

評価実験

4.1　使用するデータ

キーワードの共起語を用いるためのニュース記事として、2014 年から、2015 年を除く 2018 年までの 4 年分の日経速報記事 15639 記事を使用する。また、企業の特徴ベクトルを求めるために、CoARiJ データセットを用いる。このデータセットは、上場企業の事業内容がまとめられている。今回はそのデータセットから 2018 年の 3718 社の事業内容データを用いる。また、データの前処理として、ニュース記事と、事業内容のデータには、以下の条件を満たす語句を抽出したリストを作成している。なお、一つの文書に複数出現する語句に関しても、一つとカウントする。

- 名詞か固有名詞
- 一文字、人名や地名でない
- 助数詞可能名詞、サ変可能名詞でない
- 記事特有の語句

ニュース記事や、事業内容特有の表現を削除するようにしている。例えば、「資料」、「ご覧」、「URL」、「JPG」などの語句がここでは削除される。これらの条件から特定の名詞を抽出して、共起語の取得や、企業の特定の名詞を抽出する。また、形態素解析器には

GiNZA[9] を用いる。GiNZA では、トークンの分割単位に 3 つの「分割単位」を切り替えることができる。たとえば、「国家公務員」を分割する場合、分割単位が A だと最も短い言葉の単位を分割単位にするため、「国家/公務/員」となる。「国家公務員」を分割単位でまとめた例を表 4.1 にまとめる。今回はトークンの分割単位を B にして形態素解析を行う。分割単位を C にした場合、分割して欲しい部分が分割されない可能性がある。また、A の場合、「二酸化炭素」が「二酸化」と「炭素」に分割され、違う意味の語句が関連記事としてとれてしまうため、B を採用する。

- A: 最も短い言葉の単位
- B: A 単位 + 接尾、および一部の複合動詞
- C: 複合名詞、固有名詞、慣用句など

表 4.1　トークンの分割単位の例

分割単位	例
A	国家/ 公務/ 員
B	国家/ 公務員
C	国家公務員

4.2　実験方法

4.2.1　関連企業の定義

ここでの関連企業とは、企業の Web サイトにキーワードに関連していると判断できる材料がある企業と考える。そのため、企業の Web サイトに記述がない企業や、判断できる材料がない企業は関連企業でないと判断する。また、今回は類似度が上位 50 件に表示された企業は関連企業と判断する。評価方法の例を表 4.2 に示す。

表 4.2　評価方法の例

判断例	事例
キーワードの製造メーカー	半導体製造企業など
キーワードを用いた商品製造メーカー	半導体を用いた自動車企業など
提携先に関連企業がある	半導体製造に必要な商品メーカー
サイトにキーワードの影響がある記述	半導体不足により・・・
キーワードと関連がある語句の関連企業	半導体→ロボットやゲームなど

4.2.2　企業検索手法の評価

評価として、キーワードに、「半導体」「木材」「旅行」「音楽」「ゴム」「カメラ」の関連企業を調べる。上位 50 件の適合率と、キーワードを事業内容に含めない企業の上位 20 件の適合率を求める。適合率は、分類器の分類結果のうち正解データと一致する割合を表す。適合率の定義式を式 4.1 に示す。なお、今回の TP は類似度が上位の企業が実際に関連企業である数、FP は類似度が上位だが関連企業ではない数である。

$$\text{適合率} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{4.1}$$

4.2.3　共起語抽出の情報源の違いの影響の分析

提案手法では共起語抽出にはニュース記事を用いるが、これを企業ベクトルの構築と同様に事業内容文から共起語抽出を行う。ニュース記事から共起語を取得した場合と、事業内容から共起語を取得した場合の関連企業の評価を行う。実際に取得できた共起語と上位企業を比較を行い、ニュース記事から取得する共起語と事業内容から取得する共起語の違いについて評価していく。具体的な比較方法としては、上位 20 件の関連企業の事業内容を分類し、どのような業種があるかを確かめて比較する。

4.3 実験結果

4.3.1 企業検索結果の評価結果

4.2.2 で述べた方法で評価を行ったところ、表 4.3 に示すような結果になった。上位 50 件の適合率の平均が 0.90(270/300)、事業内容にキーワードを含まない企業だけの上位 20 件の適合率は 0.85(102/120) であった。以下でキーワードごとの結果の詳細について述べる。

表 4.3　キーワードごとの適合率の結果

キーワード	上位 50 件の適合率	含まない上位 20 件の適合率
半導体	1.00(50/50)	1.00(20/20)
木材	0.88(44/50)	0.85(17/20)
旅行	0.94(47/50)	0.90(18/20)
音楽	0.92(46/50)	0.85(17/20)
ゴム	0.72(36/50)	0.65(13/20)
カメラ	0.94(47/50)	0.85(17/20)
平均	0.90(270/300)	0.85(102/120)

キーワード「半導体」の評価

「半導体」での類似度が 0.65 以上の語句は、「シリコンウエハー」の一つだけだった。共起語を表 4.4 に示す。半導体材料のシリコンや、素子、基板、回路といった半導体用語が共起語になっている。また、電流や電圧、温度、モジュールといった語句も登場している。高性能、小型化などの現在半導体製造に欠かせない機能に関する語句も共起語になっている。

また、上位 50 件の適合率は 1.00(50/50) だった。全ての企業が「半導体」と関連のあ

る企業が上位にあがった。ほとんどの企業が半導体や、電子部品などを製造する企業だった。また、半導体製造に欠かせない商品の開発をする企業などが多かった。また、事業内容に「半導体」を含んでいない企業だけの上位 20 件の適合率も 1.00(20/20) であった。半導体メーカーだけではなく、制御装置の開発、太陽光パネルの製造や、回路基板、液晶ディスプレイ、自動車製造といった幅広い企業が上位になった。

表 4.4　「半導体」の共起語

半導体	シリコン	素子	アナログ	電流	基盤	回路
エレクトロニクス	チップ	メモリ	薄膜	電圧	小型化	原理
デバイス	技術革新	周波数	原子	高性能	マイクロ	信号
ディスプレイ	モジュール	電子	サンプル	小型	用語	センサ
温度	パワー					

キーワード「木材」の評価

「木材」での類似度が 0.65 以上の語句は、「材木」「製材」「板材」「合板」「建材」だった。共起語を表 4.5 に示す。森林、地球、といった自然に関する語句や、バイオマス、温暖化、CO2、などの環境に関する語句、住宅、建築関係の語句などが共起語になっている。

表 4.5　「木材」の共起語

木材	林業	木造	木質	森林	パルプ	バイオマス
建材	建築物	内装	チップ	グリーン	繊維	温暖化
部材	工法	世帯	機能性	燃料	資材	強度
資源	地球	プラスティック	住宅	建物	CO 2	成分
素材	活性化					

上位 50 件の適合率は 0.88(44/50) だった。多かった事業内容としては、木質バイオマス事業や、建築資材販売やリフォームなどの住宅に関する事業内容が多かった。関連企業でなかった事業内容には、投資事業、建材事業、車のシートなどの紡織の事業があっ

た。投資事業には、投資対象に木質バイオマス企業があったが、ホームページにはそれらの記述が確認できなかったため関連企業にしなかった。また、建材事業も木材を使っている製品が確認できなかった。「木材」を事業内容に含まない企業の上位 20 件での適合率は 0.85(17/20) だった。関連企業ではなかった 3 件は、先ほど関連企業とならなかった 3 件と同じ企業である。

キーワード「旅行」の評価

「旅行」での類似度が 0.65 以上の語句はなかった。共起語を表 4.6 に示す。また、GINZA の分割単位が B の場合、「旅行業」などは「旅行」と「業」に分割されない結果になっているため、共起語に「旅行」は存在していない。宿泊、航空といった旅行に欠かせない語句や、外国人、中国語などの国際に関する語句も共起語になっている。

表 4.6　「旅行」の共起語

旅行者	海外旅行	国内旅行	旅行業	旅行会社	旅行商品
旅行客	ツアー	ジェイティービー	JTB	誘客	宿泊施設
トラベル	外国人	インバウンド	観光地	航空客	航空券
もてなし	多言語	オリンピック	ホテル	中国語	空港
日本人	人気	スポット	特典	魅力	プロポーション

上位 50 件の適合率は 0.94(47/50) だった。多かった事業内容としては、ホテル事業、ツアー事業、航空会社、イベント運営などが挙げられる。関連企業でなかった事業内容は、リゾ婚、フォトウェディングの企業、新築マンションの企画・開発事業、家具やインテリアのインターネット通信販売事業である。フォトウェディングは、「旅行」という語句を事業内容に含まれていたが、旅行に関する記述が Web サイトで確認できなかった。新築マンションの規格・開発事業はホテルという語句がマッチして上位に上がる結果となっている。また、通信販売事業は、外国人に向けた日本のインテリアを販売しているため、「外国人」がマッチして上位に上がってしまったのが原因と考えられる。また、「旅行」を事業内容に含まない企業の上位 20 件の適合率は 0.90(18/20) だった。FreeWi-Fi

サービスを提供する企業、予約や翻訳をしてくれる機械やロボットを開発している企業、店舗催促をするための看板や POP の製造、旅行に関する Web アプリ開発事業などが上位に確認できた。

キーワード「音楽」の評価

「音楽」での類似度が 0.65 以上の語句はなかった。共起語を表 4.7 に示す。ラジオ、映画、ゲームといった音楽が欠かせない語句や、スピーカー、オーディオ、ディスプレイといった機器類も共起語となっている。

表 4.7 「音楽」の共起語

音楽	サウンド	映画	ストリーミング	機内
スピーカー	エアライン	オーディオ	エンターテインメント	ラジオ
ゲーム	チャンネル	毎日	ワールドコンテンツ	公式
コールセンター	イベント	映像	メディア	動画
無料	日本語	ベスト	文化	音声
オプション	ニュース	ディスプレイ	ライセンス	

上位 50 件の適合率は 0.92(46/50) だった。主な事業内容としては、広告プラットフォーム、アニメや漫画、映画などを制作する企業、テレビ関係の企業がランクインした。これらは、音楽を使用することがあるため、関連企業としている。関連企業でなかった事業では、予備校に関する企業、音声認識ソリューション、経済関係のプロモーションがあった。また、HP がない企業があったため、ここでは関連企業としていない。「コンテンツ」「映像」といった語句が認識され、上位の企業となっている。また、「音楽」を事業内容に含まない企業の上位 20 件の適合率は 0.85(17/20) だった。

キーワード「ゴム」の評価

「ゴム」での類似度が 0.65 以上の語句はなかった。共起語を表 4.8 に示す。タイヤ、トラック、ベルトなどのゴムを用いた語句があった。しかし、低燃費、お客様相談室、原材

料、産総研といった、ゴムと関連があるとは言い難い語句も存在していた。また、「横浜ゴム」や、「住友ゴム工業」など、企業名が共起語になっている。

表 4.8 「ゴム」の共起語

ゴム	横浜ゴム	住友ゴム工業	エラストマー	Tire
タイヤ	ブリヂストン	低燃費	ベルト	高分子
東洋	SUV	お客様相談室	乗用車	原材料
トラック	耐久性	産総研	部材	天然
繊維	産業技術総合研究所	樹脂	プラスティック	シート
材料	性質	石油	バス	高機能

上位 50 件の適合率は 0.72(36/50) だった。少しほかの語句よりも低い結果になった。関連企業でない事業で多かったのは、「樹脂」に関する事業が多かった。「ゴム」は自然樹脂か合成樹脂で作られるが、プラスティックなどの製品も樹脂から作られることがあるため、そのような企業が上位にきていた。関連企業で多かった事業としては、タイヤを製造する企業、ゴムホースやゴムシートなどの商品を製造する企業があった。また、事業内容に「ゴム」を含まない企業での上位 20 件の適合率は 0.65(13/20) であった。事業内容に「ゴム」を含んでいない場合は適合率が低くなっていることが分かった。

キーワード「カメラ」の評価

「カメラ」での類似度が 0.65 以上の語句は、半角カナのカメラ、「ビデオカメラ」、「デジカメ」、「キャメラ」、「マイカメラ」、「インカメラ」があった。共起語を表 4.9 に示す。ドライバーやセンサー、遠隔、道路などの運転や検知システムに関連しそうな語句や、ロボット、タブレット、端末と言った語句が共起語として抽出されていた。

上位 50 件の適合率は 0.94(47/50) だった。主な事業としては、カメラ、スマホ製造以外にも AI システムによる検知システム構築、遠隔監視といった事業や、会議アプリや QR コード読み取りアプリ事業などがあった。デジタルカメラのような製造メーカーではなく、カメラ機能を用いた製品や、カメラで取った画像を用いた事業が多かった。他に

表 4.9　「カメラ」の共起語

カメラ	デジタルカメラ	映像	解像度	センシング
ドライバー	事故	センサー	遠隔	センサ
ディスプレイ	信号	距離	動画	道路
リアルタイム	ロボット	写真	画面	次元
デジタル	精度	タブレット	音声	無線
端末	高性能	動き	セキュリティ	小型

も、カメラを製造するために必要な商品の製造もあった。また、事業内容に「カメラ」を含まない企業での上位 20 件の適合率は 0.85(17/20) であった。スマホ関連、画像ファイル保存ツール、半導体製造、画像認識を用いたシステム関連などが上位に上がっていた。

4.3.2　事業内容文の共起語取得との比較

4.2.3 で述べた方法で評価を行う。今回はキーワードを「半導体」と「木材」で比較した。

「半導体」での比較

キーワードを「半導体」にした場合の共起語比較を行う。表 4.10 に事業内容から取得した「半導体」の共起語を示す。比較的関連のありそうな語句もあるが、蘇州や海外名などあまり意味のない語句も存在していた。また、ニュース記事、事業内容から求めた関連企業の類似度上位 20 件の事業についてまとめた表 4.11 を示す。なお、半導体メーカーは、半導体そのものを製造している企業である。半導体製造関連は、半導体を製造するために必要な商品を製造する企業などがここに当たる。半導体を用いた商品は、半導体を使用して違う商品を製造している企業などがここに当たる。

ニュース記事と事業内容から共起語を求めた場合の関連企業は、半導体製造関連や半導体を用いた商品の企業が多数を占めているが、半導体メーカー以外にも幅広い事業の

企業をとることができた。

表 4.10　事業内容から求めた「半導体」の共起語

半導体	シリコン	液晶	エレクトロニクス	回路	真空
基盤	チップ	ディスプレイ	光学	応用	有機
装置	パネル	ガラス	車載	精密	モジュール
ロボット	制御	組立	イオン	プリント	蘇州
計測	部材	SINGAPORE	ユニット	EUROPE	検査

表 4.11　「半導体」上位 20 件の事業の業種の比較

業種 / 抽出方法	半導体メーカー	半導体製造関連	半導体を用いた商品	その他
ニュース記事	3	8	8	1
事業内容	4	7	8	1

「木材」での比較

キーワードを「木材」にした場合の共起語比較を行う。表 4.12 に事業内容から取得した「木材」の共起語を示す。共起語には、家具やリフォームなどの住宅に関する語句があった。また、ニュース記事、事業内容から関連企業を求めた上位 20 件を事業ごとにまとめた表 4.13 を示す。なお、建築資材は建築で必要になる木材や木材製品を製造している事業に当たる。住宅はリフォーム、解体、インテリアなどの住宅に関する事業に当たる。発電リサイクルは、木質チップやバイオマス事業などがここに当たる。複数の事業をしている場合は、代表的な事業をカウントする。

事業内容で共起語を得た場合、建築資材都住宅関係に関する事業が多いことが分かった。一方ニュース記事から関連企業を求めた場合、建築資材が 9 と多いが、その他の業種が多いなど、どの業種にも関連企業として幅広く検索結果に表示されていた。このことから、ニュース記事から共起語を求めた方が、幅広い事業を関連企業として見ることがで

きる。

表 4.12　事業内容から求めた「木材」の共起語

木材	合板	エクステリア	建材	セメント	工務店
インテリア	家具	リフォーム	倉庫	リサイクル	石油
包装	資材	加工品	建築	化成	プラスティック
原材料	合成	繊維	フィルム	内装	部材
加工	土木	ケミカル	施行	運送	ハウス

表 4.13　「木材」上位 20 件の事業の業種の比較

業種 / 抽出方法	建築資材	住宅	発電リサイクル	その他
ニュース記事	9	1	4	6
事業内容	11	7	0	2

4.4　考察

比較的適合率は高い結果になったが、「ゴム」はあまり良い結果にならなかった。適合率が良くなかった原因としては、共起語の取得がうまくいかなかったことが考えられる。「横浜ゴム」などの企業名が共起語になっていたり、「樹脂」という語句には他の語句にも関連があるため、このような結果になっていると考える。

キーワードを「木材」としたときの関連企業で、「森林」「林業」という語句のみマッチして上位に挙がった企業があった。このように、一つでも共起語とマッチすると関連企業として認識されるため、２つほど共起語とマッチしている企業が上位にくる影響が多かった。そのため、現在よりも多くの共起語をとることでどのような結果になるかを考える必要がある。また、企業によって事業内容の記述量が変わってくる。そのため、多く記述をする企業ほどマッチしやすいため上位に挙がってくる確率が高くなってしまった。

関連企業であるにもかかわらず上位に来ていなかった企業もあるため、その企業が上位に来るような工夫をする必要が考えられる。

ニュース記事から共起語を取得した場合、事業内容文から共起語を取得するよりも幅広い事業を検索結果に表示することができた。しかし、この手法では、上位にくる企業はキーワードとどのように関連があるのかが分からない。そのため、関連企業であればあるほどどのようなことがいえるのかを分かるようにする必要がある。また、ニュース記事から共起語を取得するため、多くの企業と関係しにくい語句も抽出される可能性もある。例えば、有名半導体メーカーの企業が倒産して多くのニュースになった場合、本手法では「倒産」という語句が共起語として抽出されてしまい、他の企業にあまり関係のない共起語となってしまう。

また、今回は 2018 年までのニュース記事を用いているため、最新のニュースには対応していない。「コロナ」によって影響を受ける企業を調べる際に、それ以前のニュースから関連企業を調べるため、関連企業が出てこない可能性がある。

第 5 章

結論

本論文では、企業をキーワード検索する際に、そのキーワードに直接関連していなくても、少しでも関連のある企業を検索するための手法を提案した。具体的には、キーワードから共起語を用いたクエリベクトルの構築を行い、各企業の事業内容から特徴ベクトルを構築して、コサイン類似度を求め、類似度上位を検索結果に表示する。評価実験を行い、「半導体」「木材」「旅行」「音楽」「ゴム」「カメラ」の上位 50 件の適合率が、それぞれ 1.00, 0.88, 0.94, 0.92, 0.72, 0.94 と比較的高い結果になることが確認できた。また、事業内容にキーワードを含まない企業の上位 20 件の適合率は、1, 0.85, 0.9, 0.85, 0.64, 0.85 であった。しかし、「ゴム」などの一部の語句ではあまり関連のある企業を検索結果に表示することはできなかった。また、共起語の抽出にはニュース記事を用いているが、これによって事業内容から共起語を抽出した場合に比べて幅広い企業の検索が可能になったことが実験から確認できた。

今後の課題として、ニュース記事以外のより良い共起語取得の方法や最新のニュースにも対応した関連企業検索手法の提案などが挙げられる。現在のニュース記事からの共起語取得では、関連のない語句を取得する可能性もあるため、より良い共起語を見つける方法を提案する必要がある。また、最新のニュースにも対応できるように、最新の情報を取得した検索手法も提案する必要がある。

謝辞

本論文を作成するにあたり、丁寧かつ熱心にご指導を賜りました湯本高行准教授に心よりお礼申し上げます。また、この 1 年間様々な面で関わった皆様に感謝を申し上げます。

参考文献

[1] CoARiJ データセット. https://github.com/chakki-works/CoARiJ

[2] Joulin, A. Grave, E. Bojanowski, P. Mikolov, T. Bag of tricks for efficient text classification. arXiv preprint arXiv, 1607.01759, 2016.

[3] Bojanowski, P. Grave, E. Joulin, A. Mikolov. T. Enriching word vectors with subword information. Transactions of the association for computational linguistics, Vol.5, pp.135-146, 2017.

[4] Mikolov, T. Chen, K. Corrado, G. Dean, J. Efficient estimation of word representations in vector space. arXiv preprint arXiv, 1301.3781, 2013.

[5] Agrawal, R. Imieliński, T. Swami, A. Mining association rules between sets of items in large databases. In Proceedings of the 1993 ACM SIGMOD international conference on Management of data, pp.207-216, 1993.

[6] 間瀬久雄, 徳田圭世, 森本由起子, 辻洋, 丹羽芳樹. WWW ホームページからの共起語自動抽出実験, 情報処理学会全国大会講演論文集, Vol.55, No.3, pp.72-73, 1997.

[7] 新谷研, 角田達彦, 大石巧, 長尾眞. 単語の共起頻度と出現頻度による新聞の関連記事の検索手法, 情報処理学会論文誌, Vol.38, No.4, pp.855-862, 1997.

[8] 平野正徳, 坂地泰紀, 木村笙子, 和泉潔, 松島裕康, 長尾慎太郎, 加藤惇雄. テキストマイニングを利用したテーマに関連する上場企業検索ツールの開発, 人工知能学会第二種研究会資料, Vol.2020, No.24, pp.226-233, 2020.

[9] GiNZA. https://megagonlabs.github.io/ginza/