*The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019*

# 有価証券報告書を活用した企業名の語義曖昧性解消法の一考察

## A consideration of word sense disambiguation of company name utilizing securities report

松田 裕之[*1]
Hiroyuki Matsuda

津田 和彦[*1]
Kazuhiko Tsuda

[*1] 筑波大学大学院ビジネス科学研究科
Graduate School of Business Sciences, University of Tsukuba

Word Sense Disambiguation (WSD) is a research field that distinguishes semantics from peripheral information of the target word. This research worked on WSD of company names and aimed to acquire the same classification accuracy as supervised learning with unsupervised learning. Success of WSD of company names with unsupervised learning will enable us to extract only the information of the specific business without considering the appropriate search word, without putting enormous effort on preparing teacher data. We proposed a classification method to judge word sense from similarity of word vectors by business created using securities report and word vectors of words in classification target. With the proposed method, we achieved almost the same classification accuracy as supervised learning. It is suggested that introduction of a model for determining similarity to unknown words such as fasttext, suggests that there is room for improvement of accuracy.

## 1. はじめに

従来のテキストマイニングの課題に，文字列の表記のみが集計・分析されていることがある．例えば，運転手を指す「ドライバー」と，ねじ回しを指す「ドライバー」のような多義語の課題である．このような多義語に対して，対象語の周辺情報などから語義を識別する語義曖昧性解消という研究分野がある．

本研究では，対象語を企業名に絞り，教師あり学習で構築した分類器に匹敵する分類精度を，教師データの作成を行わずに達成することを目指す．分類対象の典型例として「ヤマハ」を取り上げ，①楽器メーカーとしてのヤマハ，②二輪メーカーとしてのヤマハ，③①②のいずれでもないヤマハ，の3つの語義に識別する．

本研究によって，期待される成果は2点ある．1点目は，企業名の語義曖昧性解消ができれば，適切な検索キーワードを思いつかずとも，特定事業の情報のみを抽出することが可能となる点である．「ヤマハの二輪製造事業についてテキストデータから分析せよ」と言われれば，「ヤマハ　バイク」という条件で検索しデータを抽出する人が多いと思われるが，二輪製造事業に関するデータで「バイク」が含まれないものは多くある．一方で，適切な検索キーワードを1ユーザーが独力で全て羅列するのは不可能である．そのような場合，企業名の語義曖昧性解消手段が確立されていることは有用である．

2点目は，教師なし学習による語義曖昧性解消で，対象語の周辺情報に加え，外部知識としてシソーラスの語釈文を活用する手法が Chen ら[Chen 2014]より提案されている．このアルゴリズムを拡張し，有価証券報告書を知識抽出ソースとし，企業名の語義曖昧性解消で外部知識を活用した点である．

## 2. 語義曖昧性解消

機械学習手法による語義曖昧性解消は，一般に教師あり学習にて解決することが多い．例えば，単語「ドライバー」を含む用例を適当な数集め，各々の用例に対してその用例中の「ドライバー」の語義を付与しておく．これを教師データに，周辺語の情報などを対象語の素性として分類器を構築，識別タスクを実行する手法である．

しかし，教師あり学習による分類は教師データの作成コストが大きく，対象語が限定されてしまう問題がある．全ての単語に語義を付与する語義曖昧性解消は all-words WSD というタスクとして研究されているが，ここでは教師あり学習によるアプローチは非現実的であり，教師なし学習が用いられる．ただし，教師なし学習による識別精度は一般に教師あり学習よりも低い問題がある．

教師なし学習による語義曖昧性解消において一時 state-of-the-art の性能を示していたのが，Chen ら の手法である．Chen らは，Mikolov ら[Mikolov 2013]より提唱された Skip-gram により単語分散表現を得た後，シソーラス WordNet[wordnet]上の，多義語の語釈文中の類似単語を利用して各語義の意味ベクトルを作成，この意味ベクトルと対象語が含まれる文のコンテキストベクトルとのコサイン類似度から，語義を判定している．

本研究の提案手法は Chen らの手法を拡張したものである．Chen らは外部知識として WordNet の語釈文を活用したが，企業名について各語義を説明している情報として，本研究では有価証券報告書を活用した．

## 3. 外部知識導入による企業名の語義曖昧性解消

語義曖昧性解消の対象とする企業には「ヤマハ」を選定した．ここでいう「ヤマハ」は，ヤマハ株式会社とヤマハ発動機株式会社の2社を指す．

日経テレコンにて取得した日経新聞朝刊記事中の「ヤマハ」を含む文 2109 件を分類対象データとした．さらに，有価証券報告書上で各事業に言及する箇所を4箇所特定し，2017 年度のヤマハ株式会社およびヤマハ発動機株式会社の有価証券報告書から事業別に文を抽出し，本研究で用いる外部知識データとした．

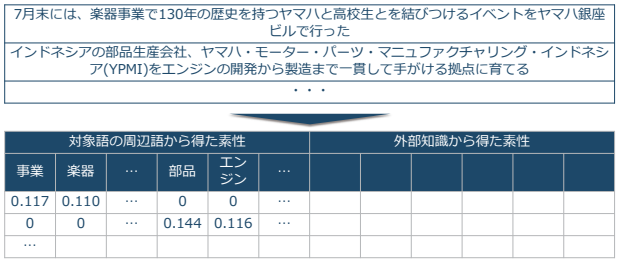
本研究では，本研究独自の手法に加え比較対象として，教師あり学習，教師なし学習，Chen らの研究をベースとした手法の3つ，計4手法で分類精度を比較した．

まず，教師あり学習/教師なし学習向け検証データの作成法を簡単に述べる．分類対象データおよび有価証券報告書から tf-idf 値が一定以上の名詞を抽出し，機械学習で活用する素性ベクトルとした．分類対象データから得たものは図 1 のようにな

連絡先：松田 裕之，筑波大学大学院ビジネス科学研究科，東京都文京区大塚 3-29-1，s1740125@s.tsukuba.ac.jp

り，これを「対象語の周辺語から得た素性ベクトル」と呼ぶ．さらに，有価証券報告書から得たものを追加すると図 2 のようになり，これを「外部知識から得た素性ベクトル」と呼ぶこととする．

*図 1. 対象語の周辺語から得た素性*



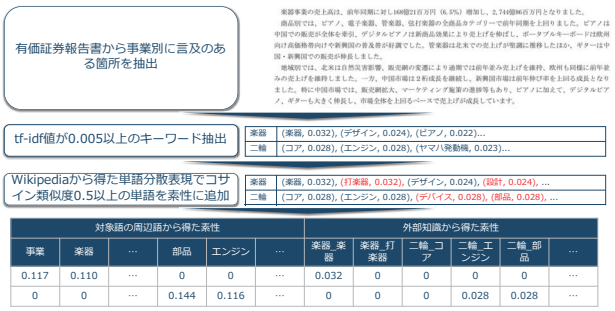




*図 2. 外部知識から得た素性*

分類方法は，教師あり学習では SVM，教師なし学習では K-means 法を実施した．

次に，Chen らの研究をベースとした手法（以下「Chen らの手法」と呼ぶ）/本研究の手法向け検証データの作成法について述べる．有価証券報告書の事業別の文から抽出した名詞について，ヤマハの事業であれば「ヤマハ」，ヤマハ発動機の事業であれば「ヤマハ発動機」の単語ベクトル（朝日新聞コーパス[田口 2017]に gensim ライブラリ[gensim]の word2vec 関数を適用して取得）とのコサイン類似度を計算した．次に，コサイン類似度の閾値$t$を 0 から 1 まで調整しながら，コサイン類似度が$t$以上の単語のみを抽出した．最後に，抽出したコサイン類似度$t$以上の単語群について単語ベクトルの平均を計算し，これを事業別単語ベクトルとみなした．

以上の事業別単語ベクトルの作成プロセスをまとめると，図 3 のようになる．

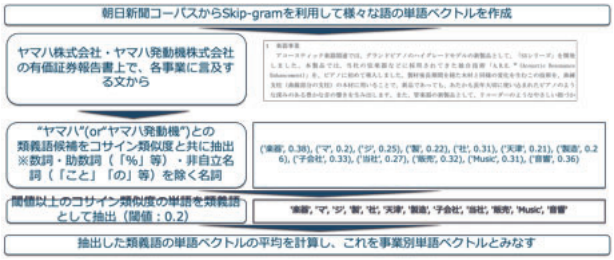



*図 3. 事業別単語ベクトルの作成*

分類方法について述べる．Chen らの手法では，まず，分類対象データに現れる名詞の単語ベクトルの平均を計算した（これを「コンテキストベクトル」と呼ぶ）．次に，事業別単語ベクトルとコンテキストベクトルのコサイン類似度を計算した．最後に，コサイン類似度の閾値$t_c$を 0 から 1 まで調整しながら，「楽器」「二輪」事業ベクトルとのコサイン類似度が$t_c$以上かつ最も高いものに分類した．複数の事業ベクトルとコサイン類似度が同値であるか，いずれの事業ベクトルともコサイン類似度が$t_c$未満である場合は，「その他」に分類した．

一方，本研究の手法では，まず，分類対象データに現れる語について各々単語ベクトルを得た．この語群に対し，コサイン類似度の閾値$t_o$を 0 から 1 まで調整しながら，事業別単語ベクトルとのコサイン類似度が$t_o$以上の語が現れる度にスコアを+1 した．以上の計算で，スコアが最も高いものに分類した．複数のスコアが同値か，いずれのスコアも 0 の場合，「その他」に分類した．

## 4. 評価結果と考察

各手法で達成した正答率は表 1 の通りである．「対象語の周辺語から得た素性ベクトル」のみを活用しているのが「外部知識活用なし」であり，「対象語の周辺語から得た素性ベクトル」に加え「外部知識から得た素性ベクトル」も活用しているのが「外部知識活用あり」である．また，Chen らの手法および本研究の手法では，コサイン類似度の閾値$t, t_c, t_o$を調整する中で正答率が最高となったときの値を示している．

*表 1. 各手法で達成した正答率*

| 分類手法 | | 正答率 |
|---|---|---|
| 教師あり学習 | SVM（外部知識活用なし） | 78% |
| | SVM（外部知識活用あり） | 80% |
| 教師なし学習 | K-means（外部知識活用なし） | 45% |
| | K-means（外部知識活用あり） | 63% |
| Chenらの手法 | 事業別単語ベクトルとコンテキストベクトルのコサイン類似度から判定 | 71% |
| 本研究の手法 | 事業別単語ベクトルと文中の語のコサイン類似度から判定 | 76% |

分類精度としては教師あり学習に匹敵する程度が求められたが，教師あり学習による正答率は 80%であったのに対し，本研究の手法による正答率は 76%と，教師あり学習には届かないものの4ポイント低いのみの水準を達成しており，目的は一定達成したと言える．

## 5. おわりに

本研究では，外部知識として有価証券報告書を活用し，事業別単語ベクトルを作成した上で，事業別単語ベクトルと分類対象中の語の類似度から分類を行う手法により，目標とする教師あり学習に近い分類精度のアルゴリズムを構築することに成功した．

今後，取り組むべき課題は主に2点挙げられる．1点目は，word2vec モデルにとっての未知語の存在が，分類精度向上の壁の1つとなった点である．fasttext モデル[Bojanowski 2016]などを活用し，未知語に対する類似性も判定することで分類精度を向上させられる可能性がある．

2点目は，アルゴリズムの汎用性の担保である．本研究は分類対象を「ヤマハ」に限定し，各種パラメータの最適化を行なっているため，分類対象を変更した場合に有効なパラメータであるかは確認できていない．分類対象を複数事業に取り組む他企業に拡張し，汎用的なアルゴリズム・パラメータであるか検証する必要がある．

## 参考文献

[Chen 2014] D.Chen, C.D.Manning: A Fast and Accurate Dependency Parser using Neural Networks, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014.

[Mikolov 2013] T.Mikolov, I.Sutskever, K.Chen, G.Corrado, J.Dean: Distributed Representations of Words and Phrases and their Compositionality, CoRR, 2013.

[wordnet] http://www.nltk.org/howto/wordnet.html, 最終アクセス日:2019-02-03

[田口 2017] 田口雄哉, 田森秀明, 人見雄太, 西鳥羽二郎, 菊田洸: 同義語を考慮した日本語の単語分散表現の学習, 情報処理学会研究報告, 2017.

[gensim] https://radimrehurek.com/gensim/index.html, 最終アクセス日:2019-02-03

[Bojanowski 2016] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov: Enriching Word Vectors with Subword Information, CoRR, 2016.