分散表現を用いたブランド名の語義曖昧性解消と小規模学習データへの応用

尾城 奈緒子[1]　竹村 彰浩[1]　白石 空[1]
[1]株式会社インテージ
{oshiro.44969, takemura-a, shiraishi-s}@intage.com

概要

従来の語義曖昧性解消で用いられてきたデータセットは、周辺単語が豊富なため辞書などの外部知識を用いた同形異義語の語義の特定が容易であることが多い。一方で、SNS で投稿される短文では、特定のブランド名であるか、判断がつかないことがある。このようなブランド名の語義曖昧性解消において、教師あり手法を適用する際の学習データの作成にかかる人的コストは軽視できない。本項では特定ブランドの外部知識と、その分散表現によって素性を構成し、小規模学習データにおいても高い精度で語義曖昧性解消を可能にする手法を紹介する。ブランド名を含む Twitter 投稿文を取得、データセットを整備し、提案手法の有効性を検証した。

1 はじめに

同形異義語とは、同じ表記で異なる意味を持つ語であり、このような多義性を持つ語の語義を特定するタスクを語義曖昧性解消と呼ぶ。SNS を用いたマーケティングでは、投稿文を分析する際にブランド名をキーワードに検索するが、対象のブランド名以外の同形異義語が含まれてしまう。例文(1)、(2)はマックで検索した場合抽出できる文である。エスティ―ローダー社メイクアップブランド「M·A·C」を対象とした場合、(1)はコスメという語が含まれるため対象であるが、(2)はポテトという語からマクドナルドの略称と推察できるため対象外と判断できる。単純な検索では、対象外の投稿文も抽出してしまうため、語義曖昧性解消が必要となる。

(1) 対象. マックのコスメ欲しい
(2) 対象外. マックのポテト食べたい

語義曖昧性解消に関する先行研究は、語義の推定に周辺単語を用いる手法[1, 2]、分散表現を用いる手法[2, 3]、知識ベースに基づく手法[2, 4]が知られている。先行研究で用いられているデータセットでの語義の特定は、辞書を確認することで人が見て明確に推定できるものが多い。一方でブランド名のように情報の改廃が頻繁に行われ、辞書への掲載が間に合わない、知識ベースの更新が追いつかない場合、辞書や WordNet などの知識ベースを用いることは困難である。また、SNS 投稿で多く見られる短文では、様々な情報が省略されるため周辺単語が少なく、テキストだけでは対象のブランド名であるか判断がつかないことがある。

教師あり手法で学習したモデルが高い精度で語義曖昧性解消が可能であることは先行研究で示されているが、学習データの作成にかかる人的コストは軽視できない。そのため、なるべく小規模の学習データで、対象の語義を特定するのが難しいブランド名の語義曖昧性解消を行う手法が求められている。本項では、小規模学習データにおいて対象のブランド名を判別する手法を提案し、SNS 投稿文に適用した結果について述べる。

2 関連研究

語義曖昧性解消では、教師あり手法においてさまざまな素性が提案されている。Raganato ら(2017)は、乱立した語義曖昧性解消において、統一の評価フレームを作成し、既存の教師あり手法と知識ベースの手法の比較実験を実施した結果、教師あり学習の中でも分散表現を用いた手法が優れていることを示した[2]。教師あり手法で、対象単語の周辺単語に加え単語の分散表現を用いて精度を上げた研究として、菅原ら(2015)の word2vec の Skip-gram モデルを用いたもの[3]、曹(2019)の BERT Multi-head attention を用いたもの[5]が挙げられる。

菅原ら(2015)では、一般的に教師あり手法を用いた語義曖昧性解消では、モデルにサポートベクトルマシン(SVM)が用いられることが多いことにも言及している。本項では、素性が学習に与える影響を検証するため、より簡易なロジスティック回帰(LR)を

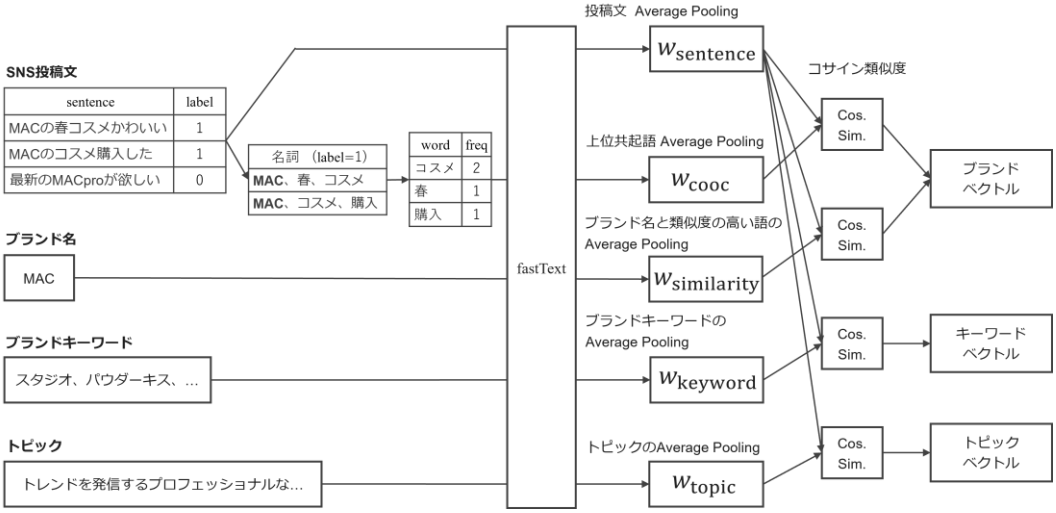

図 1　提案手法の概要

用いた。松田・津田(2019)は、企業名の語義曖昧性解消を目的として、データセットと外部知識に日経新聞の記事、有価証券報告書をそれぞれ用いて、tf-idf 値が一定以上の名詞と単語分散表現を素性とした教師あり手法の精度が高かったことを示した[6]。本研究では、松田・津田(2019)の教師あり手法を比較手法として参照した。

3 手法

本項では、企業が名付けた商品名称、ブランド名称、サービス名称を総じてブランド名として定義する。提案手法は Raganato ら(2017)にならい、分散表現を素性に組み込んで教師あり学習を行うアプローチを採用した。また、辞書未登録単語の形態素解析の結果が信頼できないことに対する対応策として、ブランドに関連すると考えられる単語と短文を外部知識として、それらの分散表現ベクトルを計算し、素性として用いた。

3.1 提案手法

図 1 に提案手法の概要を示す。本手法において用いる素性は 3 種類で構成される。1) ブランドベクトル：ブランド名そのものに関する単語の分散表現、2) キーワードベクトル：ブランドや当該カテゴリにおける関連すると考えられた単語の分散表現、3) トピックベクトル：分類に関連すると考えられた記事のタイトルの分散表現、である。

まず、分類対象の投稿文に対し、形態素解析を行い、名詞のみを抽出する。そして、抽出した名詞の平均ベクトル、三つの素性から生成されたベクトルと投稿文から生成したベクトルのコサイン類似度を最終的な素性とする。

SNS 投稿では情報が省略されるため、例えばメイクアップブランドの上位概念である化粧品カテゴリに関する直接的な情報は投稿文中に含まれないことが多い。そこで、ブランドが含まれる上位概念の単語をいくつか定義し、それらと共起する単語を用いて上位概念の特徴ベクトルを構成することとする。この特徴ベクトルを用いることで、上位概念に関連するキーワードを網羅的に指定することなく、投稿文と上位概念の類似度が計算できる。同様に、ブランドコンセプト、ブランド名からイメージされる情報、広告のキャッチコピーから構成されるトピックの特徴ベクトルとの類似度を含めることで、より効果的に曖昧性解消ができると考えられる。

分散表現には日本語学習済み fastText[1)][7]を用いた。形態素解析には MeCab[8]を利用し、辞書には mecab-ipadic-NEologd[9]を用いた。

3.2 比較手法

本項では、1) 名詞のみの形態素(以下、形態素)、2) 日本語学習済み BERT[2)]を用いたブランド名の Embedding(以下、BERT)、3) 名詞の tf-idf 値と外部情報として有価証券報告書を用いた fastText の分散表現の組み合わせ(以下、有報)、の三つの素性を比較手法とした。

1) https://fasttext.cc/docs/en/crawl-vectors.html

2) https://github.com/cl-tohoku/bert-japanese

表 1 素性として用いたキーワード・記事の一部

名称	タイプ	例
ブランド	ブランド名	MAC
キーワード	ブランドカテゴリリスト	スタジオ、ラブミー、コスメ、リップ…
トピック	記事見出・短文	トレンドを発信するプロフェッショナルな…

4 実験

4.1 データセット

検証では三つのブランド名を対象とした：1) エスティローダー社メイクアップブランド「M·A·C」、2) カネボウ化粧品メイクアップブランド「KATE」、3) 日本コカ・コーラ社コーヒーブランドの「ジョージア」である。それぞれ対象外の語義として 1) マクドナルドの略称やアップル社のコンピュータブランド、2) 人名や服飾ブランド、3) 国名や州名、が挙げられる。

ブランド名を含む、2020 年 4 月～2022 年 3 月の Twitter 投稿文（引用・返信・リツイート）を 1 ブランドにつき 2 万件、合計 6 万件抽出し、投稿ごとに、それが対象のブランド名を含むかアノテーションを実施した。まず、投稿文から容易に類推できる範囲で対象のブランド名であるかどうかの判断を行い、投稿文のみでは判断が難しいと考えられた場合は、当該ユーザの前後の投稿やプロフィール欄といった、SNS 特有の周辺情報を見て判断を行った。周辺情報確認後も判断がつかなかった場合は対象外とした。それぞれの対象ブランドについて、提案手法で使用するキーワード・記事を公式ホームページ、公式 SNS、Wikipedia から取得した。表 1 に例として検討した素性のキーワード・トピックの一例を示す。

4.2 実験設定

小規模学習データでの提案手法の効果を測るため、提案手法 4 パターンと比較手法 3 種の合計 7 種すべての素性を用いて、学習データ 100 件をランダム抽出し、対象・対象外の 2 値判別を実施した。また、学習データ件数増加による精度への影響を確認するため、学習データ数を 100 件から 3,000 件まで 100 件刻みで増加させていった際の精度の比較を実施した。2 値判別を実施するにあたり、対象が含まれる割合が低いもの(M·A·C:6%)、中程度のもの(KATE:26%)、高いもの(ジョージア:54%)をデータセットとして用意した。語義曖昧性解消で用いられるデータセットである SemEval-2010 Japanese WSD Task データセット[10]では、学習データとして単語あたり例文が 50 文用意されるのに対し、100 件あたりの出現数では「M·A·C」(6 件)と「KATE」(26 件)は少ない。また、用意したデータセットは不均衡のものが含まれるため、評価には F1-score を用いた。

表 2　素性とモデル精度(学習データ 100 件)
Br.:ブランドベクトル、Ke.:キーワードベクトル、To.:トピックベクトル、M: M·A·C、K:KATE、G:ジョージア

素性			モデル	F1-score		
Br.	**Ke.**	**To.**		**M**	**K**	**G**
✓			LR	.63	.88	.96
✓	✓		LR	.69	.90	.96
✓		✓	LR	.69	.90	**.97**
✓	✓	✓	LR	**.71**	**.91**	.96
形態素			LR	.15	.74	.94
BERT			LightGBM	.27	.85	.96
有報			SVM	.17	.85	.95

4.3 実験結果

4.3.1 提案手法の有効性検証

表 2 に学習データ 100 件での各モデルの F1-score を示す。まず提案手法と比較手法の比較より、ブランドベクトルのみの素性でどの比較手法よりも高い精度を得ることができた。また、提案手法の素性の比較より、ブランドベクトルにキーワードベクトル、トピックベクトルのどちらを追加しても精度が向上しているため、ブランドカテゴリとトピックの情報が有用な特徴であるといえる。二つのベクトルを追加した提案手法は、不均衡データである「M·A·C」、「KATE」に関しては精度が向上したが、「ジョージア」に関してはトピックベクトルのみの情報を用いた方が、精度が高いことがわかった。小規模学習データでも偏りの少ない場合は、ブランドベクトルよりもトピックベクトルの方が有効な情報であるといえる。

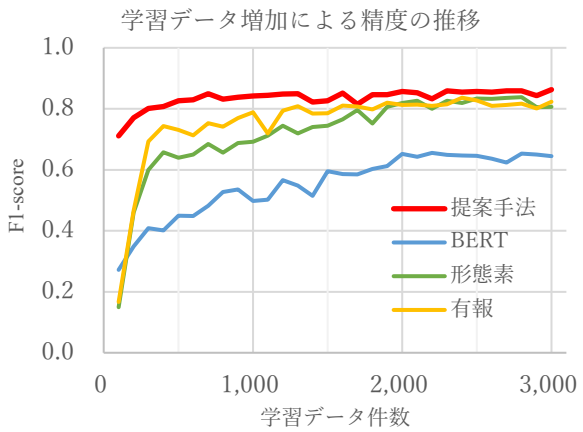

図 2 「M·A·C」学習データ件数と精度の推移

4.3.2 学習データ件数増加による精度検証

「M·A·C」の学習データ件数増加による精度を図 2 に示す。他二つのブランドは付録に示した。図 2 より、比較手法は学習データ 500 件まで精度が急激に上がっているが、提案手法は学習データ 100 件から精度の上昇は緩やかである。また、この結果は不均衡データである「M·A·C」と「KATE」のときに顕著であったことからも、提案手法は不均衡かつ小規模学習データに対して有効な素性であるといえる。

提案手法と形態素を比較すると、学習データ 3,000 件まではどのブランドでも提案手法の精度が高かった。小規模学習データでは出現する単語の豊富さの影響をより強く受けていると考えられる。有報を素性とした場合、形態素に比べて精度向上が見込めるが、提案手法の方が多くのケースで精度が高かった。ブランド名の語義特定においては、有価証券報告書の情報よりも、ブランドカテゴリやイメージワードから類似度を計算した素性のほうが有効であると考えられる。素性に BERT を用いた結果は、不均衡かつ小規模学習データの場合、どの手法よりも大きく劣ったが、均衡なデータで学習データ件数を増加させると、提案手法よりも精度が高くなった(付録図 3：「ジョージア」)。

5 考察

4.3.1 項の実験結果より、学習データが 100 件程度の小規模データでは、提案手法の有効性が確認できた。小規模学習データでも判別が上手くいく要因として、SNS 投稿文は、事業内容よりもブランドカテゴリやコンセプト・イメージに関連する投稿が多いためと考えられる。提案手法は、特に対象出現率が低い「M·A·C」データに対して有効であったため、ブランド・トピックのブランド名と類似したカテゴリやイメージは、学習データに情報が足りない場合に補える素性であるといえる。

ブランド名の語義曖昧性について本項で検証した範囲では、ブランドコンセプト・イメージ・想定シーンといった素性の有効性も確認できた。本項では、外部情報としてブランドに関連するキーワード・トピックを含めることで、精度の向上が見込めることは検証できたが、キーワード・トピックは主観で選定しており、客観的に統一した条件や検証を行っていない。

6 まとめ

本項では、語義曖昧性の解消の手法を用いて、SNS 投稿文のようにラベルが明確でない小規模学習データに対して対象のブランド名を判別するモデルを提案した。提案手法は、特に学習データが少ない場合に、既存手法より高い精度を出すことが可能である。特にブランド名の語義を推定する上で有効であったのは、外部情報として追加したブランドカテゴリ、コンセプトやイメージを反映するトピックであることを示した。外部情報として指定するキーワードの客観的な条件を指定し、既存研究で用いられているモデルを用いた精度検証を今後の課題とする。

謝辞

本研究は株式会社インテージホールディングスグループ R&D センターの助成を受けている。データセット作成にご協力くださった株式会社インテージの高崎綾子さん、岩元くるみさんに深く感謝申し上げます。

参考文献

[1] David Yarowsky. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. In ACL, pp. 189–196, 1995.

[2] Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. Word Sense Disambiguation: A Unified Evaluation Framework and Empirical Comparison. In EACL, Vol1, pp. 99–110, 2017.

[3] 菅原拓夢，笹野遼平，高村大也，奥村学，単語の分散表現を用いた語義曖昧性解消，言語処理学会第21 回年次大会発表論文集，pp648-651，2015.
[4] 松田耕史，高村大也，奥村学. 知識ベースに基づいた語義曖昧性解消における教師データの活用. 第26 回人工知能学会全国大会論文集， 2012
[5]曹鋭，田中裕隆，白静，馬ブン，新納浩幸. BERT を利用した教師あり学習による語義曖昧性解消. 言語資源活用ワークショップ発表論文集，vol. 4, pp.273-279, 2019.
[6] 松田裕之， 津田和彦. 有価証券報告書を活用した企業名の語義曖昧性解消の一考察. 第 33 回人工知能学会全国大会論文集, 2019.
[7] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759, 2016.
[8] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying Conditional Random Fields to Japanese Morphological Analysis, In EMNLP, pp.230-237, 2004.
[9] 佐藤敏紀，橋本泰一，奥村学.単語分かち書きシステム NEologd の運用--文書分類を例にして. 情報処理学会自然言語処理研究会研究報告. Vol. NL-229, no. 15, pp.1-14, 2016.
[10] Manabu Okumura, Kiyoaki Shirai., Kanako Komiya., and Hikaru Yokono. SemEval-2010 task: Japanese WSD. In: SemEval, pp. 69–74, 2010

A 付録

4.3.2 項の「KATE」、「ジョージア」における学習データ件数が増加した際の精度の推移を図 3、4 に示す。

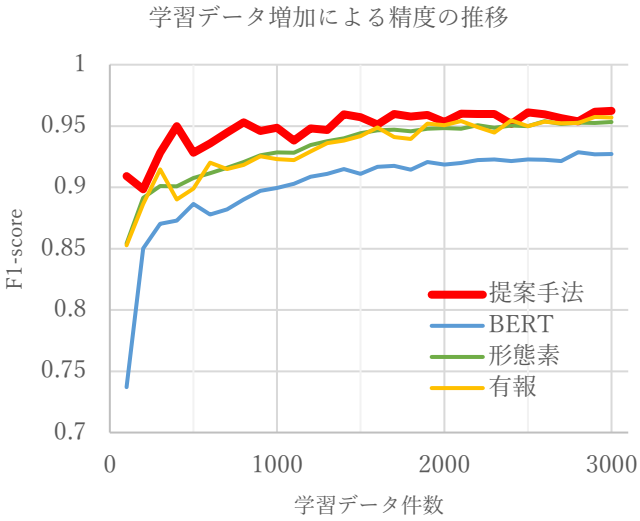

図 3　「KATE」の学習データ件数と精度の推移

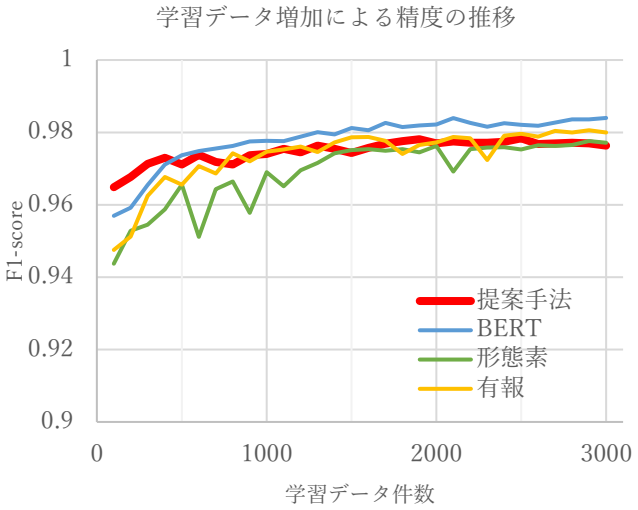

図 4　「ジョージア」の学習データ件数と精度の推移