Bilingual Word Embeddings によるターゲット言語の教師データを必要としない感情分析

荘司 響之介　　新納 浩幸　　古宮 嘉那子
茨城大学 工学部 情報工学科

{15t4034w, hiroyuki.shinnou.0828, kanako.komiya,nlp}@vc.ibaraki.ac.jp

1　はじめに

本論文では Bilingual Word Embeddings　(以下、BWE) を用いることで、教師データを利用しない感情分析を試みる。

感情分析とはレビュー文書が肯定的なものか、否定的なものかを判定するタスクである。これは文書分類の一種であり、教師あり学習を用いて解決できる。しかし教師あり学習には大量のラベル付きデータ（教師データ）が必要であり、このデータの構築コストが高いという問題がある。ただし英語などのメジャーな言語に対しては、ラベル付けされたデータが既に存在していることも多い。この場合、英語側では分類器を学習できるため、その学習できた知識を、タスクの対象となっている言語側へ転移できれば、ターゲット言語での教師データを利用せずに、分類器を構築できる。本論文はそのような転移を行うために BWE を利用する。

BWE とは英語や日本語などの異なる言語の分散表現を共通に扱う枠組みである。例えば英語 “dog” の分散表現と日本語「犬」の分散表現は、学習基のコーパスが異なるために、異なるベクトルであるが、概念としては同じなので、同一のベクトルとして表現できるはずである。このようなアイデアのもと、異なる言語の分散表現を同一したもの、あるいはそれらの変換を実現したものが BWE である。

本論文では英語のラベル付き文書を BWE を用いてベクトル化し、そのベクトルを基に分類器を学習する。次にターゲット領域の文書となる日本語文書を BWE を用いてベクトル化し、先の分類器によって識別する。これによってターゲット領域側のラベル付き文書を全く利用せずに、感情分析が可能となる。

実験では、提案手法（BWE を用いて日本語の教師データを用いずに日本語の感情分析を行う手法）と比較するために、日本語のテスト文書を英語に自動翻訳することで英語側で作った分類器を利用する手法を試した。

2　関連研究

本研究は BWE の応用と位置づけられる。この観点から関連研究を述べる。

まず、BWE の構築方法について簡潔に述べる。BWE の構築には４つのアプローチが用いられる。

１つ目は、モノリンガルマッピングである。このアプローチでは、単一言語の分散表現をその言語のコーパスから作成し、違う言語同士の分散表現を用いて線形変換を学習する。これにより、未知の単語をソース言語からターゲット言語にマッピングできる [5]。２つ目は、疑似クロスリンガルである。このアプローチは、まず違う言語が混在したコーパスを作成し、既存の分散表現モデルをそのコーパス上で学習するという手法である [6]。３つ目は、クロスリンガルである。パラレルコーパス上でそれぞれの分散表現を学習すると同時に、異なる言語間の制約を最適化することで、似た意味の単語の分散表現を共有空間上で近づけることができる [2]。４つ目は、ジョイントオプティマイゼーションである。このアプローチでは、異なる言語間の最適化に連帯して、単一言語の組み合わせ最適化とクロスリンガルの損失関数の最適化を行う [3]。

次に、BWE の応用について述べる。基本的に BWE は翻訳システムに利用できる [7]。翻訳以外の利用も可能である。例えば Chenggang Mi らはウイガル語の借用語を識別するために BWE を利用している [4]。具体的には借用語の候補リストを生成するために BWE を利用している。まず、ソース言語であるウイガル語のコーパスとターゲット言語であるドナー言語（中国語、アラビア語、ペルシャ語、ロシア語）のコーパスを用い Closslingual Word Embeddings(以下 CWE)

を学習する。そして、ウイガル語のコーパスから取り出した単語とドナー言語のコーパスから取り出した単語間の距離を構築した CWE から求め、その距離がスレッショルド ρ より小さいもの同士を取り出していくことでリストを生成する。

3 提案手法

ここではまず英語と日本語の BWE を構築する。次に構築した BWE を用いて、英語の教師データ（ラベル付き文書）を BWE の列で表現する。次にこの BWE の列をベクトル化する。これによって分類器を学習する。識別では日本語のテストデータ（ラベルなし文書）を BWE の列で表現し、英語の場合と同様に、それをベクトル化する。その文書ベクトルを先に構築した分類器を用いて識別する (図 1 参照)。

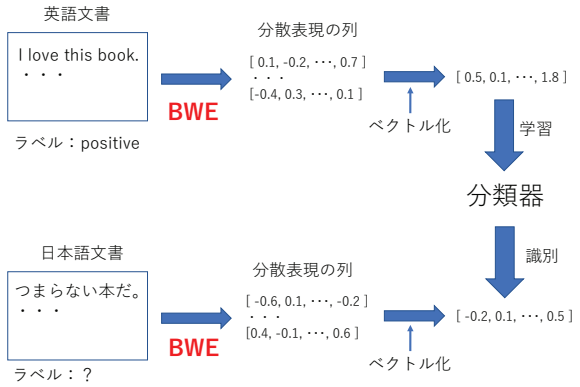

図 1: 提案手法の概要図

以下、BWE の構築とベクトル化の詳細を述べる。

3.1 BWE の構築

本論文では VecMap [1] を利用して、英語と日本語の BWE を構築する。

VecMap は BWE を学習するためのフレームワークを実装したプログラムである [1]。BWE を構築するスクリプトや単語の翻訳、類似性/関連性、類推を評価するツールを内包している。

VecMap は以下のように使用する。

まず、word2vec などを利用してそれぞれの言語の分散表現を作成する。そして、それらを用いて BWE を構築する。その際、以下の 4 つのモードの中から 1 つを選択する。

Supervised
大きい訓練辞書を持っている場合に推奨

Semi-supervised
小さい seed dictionary を持っている場合に推奨

Identical
seed dictionary を持っておらず、多義語を区別しない場合に推奨

Unsupervised()
seed dictionary を持っておらず、多義語を区別する場合に推奨

ここでは以下の手順により BWE を作成した。

1. 日本語の分散表現の作成

   gensim を用いて日本語の分散表現を作成した。コーパスは日本語の Wikipedia のデータを使用した [2]。

2. 英語の分散表現の作成

   gensim を用いて英語の分散表現を作成した。コーパスは英語の Wikipedia のデータを使用した [3]。

3. VecMap による BWE の作成

   VecMap を以下のように実行して、日本語と英語の BWE を作成した。

```
> python3 map_embeddings.py
        --semi_supervised
          3000_common_words.en2ja
          wiki_en.emb
          wiki_ja.emb
          wiki_en_semi.bwe
          wiki_ja_semi.bwe
```

   モードは Semi-supervised を利用した。また `3000_common_words.en2ja` は seed dictionary 、`wiki_en.emb` は英語の分散表現、 `wiki_ja.emb` は日本語の分散表現、`wiki_en_semi.bwe` は英語の BWE 、`wiki_ja_semi.bwe` は日本語の BWE である。

[1] https://github.com/artetxem/vecmap

[2] https://dumps.wikimedia.org/jawiki/latest/jawiki-latest-pages-articles.xml.bz2

[3] https://dumps.wikimedia.org/enwiki/latest/enwiki-latest-pages-articles.xml.bz2

3.2 BWE による文書のベクトル化

作成した BWE を用いて以下の手法で文書をベクトル化する。

$$S = \frac{w_1 + w_2 + ... + w_i}{i}$$

ここで、S は文のベクトル、wi は i 番目の単語のベクトルを表している。

4 実験

4.1 英日の感情分析データ

BWE を利用した感情分析の実験のために、以下のサイトから英語と日本語の感情分析のデータをダウンロードした。

`https://webis.de/data/webis-cls-10.html`

このデータは、日本語と英語でそれぞれ books、DVD、music の 3 つの領域を持ち、各データ（文書）数は 2000 である。また、日本語のテスト文書を英語に翻訳した文書を持つ。

4.2 翻訳を利用した感情分析

BWE による文書のベクトル化と同様に、英語の分散表現を用いて英語の訓練文書と日本語から英語に翻訳したテスト文書をベクトル化した。比較のため、英語の訓練文書を用いて学習した分類器を利用して、英語のテスト文書の感情分析を行った。学習アルゴリズムには scikit-learn の SVM を用い、C パラメータは 10 で固定した。結果を表 1 に示す。

表 1: 翻訳による感情分析（英語の分散表現）

テスト文書	books	DVD	music
日本語テスト文書を日英翻訳した文書	**0.69**	0.70	**0.72**
英語テスト文書	**0.77**	0.76	**0.78**

さらに、 Bag of Words (以下 BoW) を用いて翻訳を利用した英日の感情分析も行った。こちらも比較のため英英の感情分析も行った。学習アルゴリズムには scikit-learn の SVM を用い、C パラメータは 100 で固定した。結果を表 2 に示す。

表 2: 翻訳による感情分析（BoW を用いた場合）

テスト文書	books	DVD	music
日本語テスト文書を日英翻訳した文書	0.66	**0.79**	0.70
英語テスト文書	**0.77**	**0.77**	0.74

4.3 BWE を利用した感情分析

BWE を用いて文書をベクトル化し、英日の感情分析を行った。学習アルゴリズムには scikit-learn の SVM を用いた。C パラメータは 1000 で固定した。結果を表 3 に示す。

表 3: BWE による感情分析

テスト文書	books	DVD	music
日本語テスト文書	0.64	0.69	0.70
英語テスト文書	0.75	0.74	0.76

5 考察

まず、表 1 と表 2 を比較すると、結果があまり変わらないことがわかる。つまり、感情分析において、文のベクトル化の手法として単語の分散表現を用いる場合と BOW を用いる場合とでは精度の違いはほとんどないと言える。しかし、 BOW を用いる場合、ベクトルの特徴量は単語数に依存するため、単語数が多い場合は時間がかかってしまう。そのような場合は、分散表現を用いるのが良いだろう。

次に、表 1 から、翻訳を利用した英日の感情分析と英語の分散表現を利用した英英の感情分析を比較すると、すべての領域において後者の方が明らかに正解率が高いことがわかる。つまり、英語の文書を用いて分類器を作成した場合、翻訳された英語の文書よりも、もともと英語で書かれた文書を分類する方が精度は高くなるということが言える。分類機の学習に用いた文書とベクトル化の方法は同じであるため、このような結果になる主な理由は、翻訳にあるのは確実である。おそらく、日本語にはあるが英語にはない表現が原因だろう。

例えば、「いただきます」や「ごちそうさま」を表す英語はない。もし、無理やり翻訳しようとすれば、”Let’s eat.(さあ食べましょう)” や ”I’m full.(満腹です)” となる。「いただきます」や「ごちそうさま」は相手を気遣い感謝を伝えるという意味合いが強いため、これらの訳は実際の意味合いとは異なっている。感情を表す表現としては、「もどかしい」などがある。これを翻訳しようとすると ”irritating(イライラする)” や ”be impatient(我慢できない)” など、「イライラする」といった意味合いになる。しかし、「もどかしい」は必ずしもそれに近い感情とは限らない。これらのような、英語で表現できない日本語の存在が精度を下げていると考えられる。

次に、表 1 と表 3 の結果を比べると、すべての領域において、 BWE を利用した英日の感情分析は、翻訳を利用した場合よりも精度が少し低いことがわかる。前述した英語で表現できない日本語も原因の一つであると考えられるが、同じ意味の単語でも文脈によって意味合いにわずかな違いが生じることも原因だろう。

例えば、「カナダの首都はどこですか?」という文を英語に翻訳するとする。もし、「カナダ東部」という回答を期待するなら ”Where is the capital of Canada?” と翻訳されるが、「オタワ」という回答を期待する場合、 ”What is the capital of Canada?” と翻訳される。単語だけで見れば、通常「どこ」は ”where” と訳されるが、文の意味によっては ”what” と訳されることもある。このように、文の意味合いによって単語の訳し方も変わってくる。

しかし、今回は文をベクトル化する際、単純に単語ベクトルの平均をとる方法を用いたため、文ごとの単語の意味合いの違いを表現することができなかったと考えられる。

6 おわりに

本論文では BWE を用いることで、教師データを利用しない感情分析を試みた。BWE を利用することで英語などのメジャーな言語側で分類器を学習する。つまり、ターゲット領域（日本語）の教師データ無しで分類機を学習する。

実験では、翻訳を用いた感情分析と比較することで、提案手法の優位性を示した。しかし、翻訳する手間がかからないという意味では提案手法が優位であるが、精度はわずかに劣る結果となった。これは、文をベクトルする際、文脈による単語の意味合いの違いを表現できなかったためだと考えられる。ある単語の前後の単語にも注目することでその問題を解決できる可能性があるため、今後はその方向で研究を進めたい。

参考文献

[1] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, 2018.

[2] Karl Moritz Hermann and Phil Blunsom. Multilingual distributed representations without word alignment. *arXiv preprint arXiv:1312.6173*, 2013.

[3] Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. Inducing crosslingual distributed representations of words. *Proceedings of COLING 2012*, pp. 1459–1474, 2012.

[4] Chenggang Mi, Yating Yang, Lei Wang, Xi Zhou, and Tonghai Jiang. Toward better loanword identification in uyghur using cross-lingual word embeddings. In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 3027–3037, 2018.

[5] Tomas Mikolov, Quoc V Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*, 2013.

[6] Min Xiao and Yuhong Guo. Distributed word representation learning for cross-lingual dependency parsing. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pp. 119–129, 2014.

[7] Will Y Zou, Richard Socher, Daniel Cer, and Christopher D Manning. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1393–1398, 2013.