Bilingual Word Embeddingsによる『岩波国語辞典』の語義と『分類語彙表』の語義の対応付け

平林 照雄　　古宮 嘉那子　　新納 浩幸
茨城大学大学院理工学研究科情報工学専攻
{18nm736g, kanako.komiya.nlp, hiroyuki.shinnou.0828}@vc.ibaraki.ac.jp

1　はじめに

近年、同一のコーパス上に複数のタグセットによるタグが付与されているケースが増加している。例えば、『現代日本語書き言葉コーパス』[3] では、『岩波国語辞典』によるタグ付与が行われた後、『分類語彙表』によるタグ付与が行われている。これらのタグは参照している辞書が異なることから、異なる語義の付与が行われているが、共に文章中の単語に一意の意味を付与しているという点で関連性がある。

本研究では、『岩波国語辞典』の語義タグが付与された文章と、『分類語彙表』の分類番号が付与された文章からそれぞれ分散表現を作成し、語義タグと分類番号の対応をとる際に Bilingual Word Embeddings を利用し、その有用性を調査する。

2　関連研究

Bilingual Word Embeddings(以下 BWE と略す) は、 主に言語間を横断する分散表現のモデルの作成のアプローチから 4 種類に分けられる。一つ目は単一言語マッピングである。あらかじめ単一言語で学習済みの分散表現を、別の言語空間に投射する変換行列を学習する手法である。Mikolov ら [4] は、言語間に成立する幾何学的関係が言語間で類似していることを主張し、変換行列を用いた線形射影によってある言語のベクトル空間を別の言語の空間に変換することが可能であることを示唆した。また特定の言語をほとんど想定していないため、異なる言語同士の単語ペアや翻訳テーブルの拡張、改良に寄与できるとしている。二つ目は擬似クロスリンガルである。異なる言語の文脈を混在させたコーパスを作成し、そのコーパスに対し既存の単語分散表現モデルを適用する手法である。Xiao と Guo ら [5] は、翻訳ペアを活用する最初の疑似クロスリンガル方式を提案した。彼らはウィクショナリーを用いてソース言語コーパスの単語をすべてターゲット言語に翻訳し、ノイズを除去してから各翻訳ペアが同一の分散表現を持つようにプレースホルダに置き換えて学習する研究を行った。三つ目はクロスリンガルである。パラレルコーパスに対して分散表現を求め、異なる言語間の制約を最適化し、共有ベクトル空間内での類義語の分散表現がお互いに近くなるようにする方法である。Hermann と Blunsom ら [1] は、それぞれの言語で書かれた文章を分散表現化するモデルの出力に、最小二乗法を用いることで学習する研究を行った。四つ目は joint optimization である。クロスリンガルでの制約だけではなく、単言語またはクロスリンガルの目標を同時に最適化する手法である。Klementiev ら [2] は、joint optimization の手法を初めて行った。

BWE の応用研究として、Zou ら [6] は、Matrix Factorizartion を用いて、クロスリンガルの目標を最適化する joint optimization の手法により BWE を構築し、機械翻訳に応用している。本研究は、単一言語マッピングの手法により BWE を構築し、語義の対応に応用していると位置づけられる。我々が知る限り、本論文は BWE を語義の対応に応用した初めての論文である。

3　提案手法

本研究では、コーパス『現代日本語書き言葉コーパス』[3] から、同一の文章に『岩波国語辞典』の語義タグ (以下語義タグと略す) を付与し作成した分かち書き文から作成した分散表現と、『分類語彙表』の分類番号（以下分類番号と略す）を付与し作成した分かち書き文から作成した分散表現の間に BWE を適用し、語義タグと分類番号の対応をとる。

3.1 『岩波国語辞典』における語義タグ

『岩波国語辞典』では、単語の語義を特定するために “21128-0-0-1-0” のような語義タグが付与されている。これは [見出し ID]-[複合語 ID]-[大分類 ID]-[中分類 ID]-[小分類 ID] を並べたもので、分類に対応がないとき ID の値は 0 である。例えば、「しじょう【市場】」という単語は、複合語を無視すると、表 1 のような語義を持ち、それぞれ語義タグが付与されている。

表 1: 『岩波国語辞典』における「市場 (しじょう)」

語義タグ	意味
21128-0-0-1-0	<1> いちば。「青果—」
21128-0-0-2-0	<2> 売行き先。「—が広い」
21128-0-0-3-0	<3> 〔経済〕売手と買手とが規則的に出会って取引を行う組織。「証券—」

3.2 『分類語彙表』

『分類語彙表』とは、語を意味によって分類・整理したシソーラス（類義語集）である。[1] 一つのレコードは「レコード ID 番号／見出し番号／レコード種別／類／部門／中項目／分類項目／分類番号／段落番号／小段落番号／語番号／見出し／見出し本体／読み／逆読み」という要素から構成される。分類番号は「類／部門／中項目／分類項目」を表す 5 桁からなる数字である。例えば「市場」という言葉は、『分類語彙表』では 2 箇所に登録されている多義語である。それぞれの分類番号は 1,2600、1,2640 であり、表 2 のように分類されている。

表 2: 『分類語彙表』における「市場」

分類番号	類	部門	中項目	分類項目
1.2600	体	主体	社会	社会・世界
1.2640	体	主体	社会	事務所・市場・駅など

このように、語義タグと分類番号は粒度が異なる。「市場」を例にとると、“21128-0-0-1-0” と “1.2640” が対応し、“21128-0-0-2-0”,“21128-0-0-3-0” と “1.2600” が対応すると考えられる。

また本研究では、語義タグと分類番号は『現代日本語書き言葉コーパス』に付与されている物のみを用い、各タグの参照元である辞書を研究には用いない。

[1] http://pj.ninjal.ac.jp/corpus_center/goihyo.html

3.3 Bilingual Word Embeddings

ここでは、BWE の手法のうち、単一言語マッピングについて説明する。

単一言語マッピングは、二言語でそれぞれ単語の分散表現を作成し、二言語間で意味が似ている分散表現を近づけるように共通ベクトル空間にマッピングすることで、二言語の対応をとるという手法である。また、言語の間に成立する幾何学的関係が言語間で類似していることから、線形射影 W によってある言語のベクトル空間を別の言語の空間に変換することが可能であることが示されている。本研究では、最も簡易な線形射影 W を学習した。

4 実験

4.1 実験設定

本研究は、以下の 3 ステップにより実験を行う。

1. コーパスを語義タグの分かち書きに変換したものと分類番号の分かち書きに変換したものをそれぞれ作成し、word2vec [2] で学習し分散表現を作成する。
2. 1. で作成した分散表現のうち単義語の普通名詞を用いて、語義タグから、分類番号への線形変換 W を学習する。
3. 多義語の持つ、語義タグの分散表現に、2. で学習した W を適用して分類番号のベクトル空間に線形射影を行った分散表現と、多義語がとりうる分類番号の分散表現との cos 類似度をそれぞれ求め、最も高い分類番号を語義タグと対応しているとみなし、正解率を求める。

本研究で用いるコーパス『現代日本語書き言葉コーパス』は、単語のべ数約 140,000、単語異なり数 25,321 で、語義タグによる分かち書きに変換した時、単語異なり数 26,713 となり、分類番号による分かち書きに変換した時、単語異なり数 3,164 となる。word2vec は、アルゴリズムは C-BoW を利用し、次元数を 200、ウィンドウ幅を 5、反復回数を 5、バッチサイズを 1000、min-count を 1 として学習を行った。

本研究における「単義語」とは、『岩波国語辞典』において二つ以上の語義を持たず、『分類語彙表』によっ

[2] https://code.google.com/archive/p/word2vec/

て分類番号が付与されていない単語をさし、単義語の普通名詞は 104 単語存在した。W は 200 次元× 200 次元の線形変換で、損失関数の最適化のアルゴリズムを Adam とし、反復回数を 1015 として学習した。

本研究では、多義語として文章中に 50 回以上出現する名詞、「関係」「技術」「現場」「子供」「時間」「市場」「電話」「場所」「前」の 9 単語、25 語義を対象とした。また、各語義タグに対して正解の分類番号は、コーパス中に付与された各語義タグに最も付与された分類番号とした。

比較実験として単語に最も付与された『分類語彙表』の分類番号を、単語の各語義の分類番号とした場合 (最多出現語義の付与) の正解率も求めた。

4.2 実験結果

対象の 9 単語 25 語義に、本研究により推定された分類番号と、比較実験 (最多出現語義) により推定された分類番号の一覧は表 3 である。正解の語義を太字にした。「前」の語義の「X-X-X-X」とは、新語義を意味し、本研究においては、語義の一つとみなし実験を行った。本研究では、9 単語 25 語義に対して、正解の語義と対応が取れた語義は 15 語義存在し正解率は 60.0%であった。比較実験 (最多出現語義) では、9 単語 25 語義に対して、正解の語義と対応が取れた語義は 16 語義存在し正解率は 64.0%であった。またランダムに語義を付与したときの正解率は 41.5%となる。

4.3 考察

本研究の正解率は、比較実験 (最多出現語義) の正解率と比較しわずかに低くなっている。しかし、既に述べたように語義タグと分類番号は必ずしも 1:1 の対応関係があるわけではない。本研究では、語義タグの分散表現を分類番号のベクトル空間にマッピングをすることで、分類番号を構成する語義タグの割合を調べることが出来るという長所があると考えられる。

また本実験のタスクは、学習を行ったコーパスと正解率を求めるコーパスが同じ範囲であり、なおかつ学習段階で学習した語義タグと分類番号のペアを正解としているから、実験結果が良くなり易いと予想される。これを避けるには範囲の問題では学習データとテストデータを分ける必要があり、また学習した語義タグと分類番号のペアを他の指標で評価する必要がある。

表 3: 『岩波国語辞典』と『分類語彙表』の対応表

単語	語幹	語義	分類番号		
			本研究	正解	比較
関係	9667	0-0-1-0	**1.1110**	1.1110	**1.1110**
		0-0-2-0	**1.1110**	1.1110	**1.1110**
		0-0-3-0	**1.1110**	1.1110	**1.1110**
技術	10703	0-0-1-0	**1.3850**	1.3850	**1.3850**
		0-0-2-0	**1.3421**	1.3421	1.3850
現場	15615	0-0-1-0	**1.1700**	1.1700	1.2620
		0-0-2-0	1.1700	1.2620	**1.2620**
子供	17877	0-0-1-0	**1.2050**	1.2050	**1.2050**
		0-0-2-0	1.2050	1.2130	1.2050
時間	20676	0-0-1-0	1.1962	1.1600	**1.1600**
		0-0-2-0	**1.1962**	1.1962	1.1600
		0-0-3-0	**1.1600**	1.1600	**1.1600**
		0-0-4-0	1.1962	1.1600	**1.1600**
市場	21128	0-0-1-0	1.2600	1.2640	1.2600
		0-0-2-0	**1.2600**	1.2600	**1.2600**
		0-0-3-0	**1.2600**	1.2600	**1.2600**
電話	35881	0-0-1-0	**1.3122**	1.3122	**1.3122**
		0-0-2-0	**1.4620**	1.4620	1.3122
場所	41150	0-0-1-0	**1.1700**	1.1700	**1.1700**
		0-0-2-0	1.3833	1.1700	**1.1700**
前	48488	0-0-1-1	3.1670	1.1740	1.1670
		0-0-2-0	**1.1740**	1.1740	1.1670
		0-0-2-1	3.1670	1.1670	**1.1670**
		0-0-2-2	1.1740	1.1670	**1.1670**
		X-X-X-X	1.1740	1.1635	1.1670

さらに、語義の選択を行う際に、語義タグの分散表現が BWE により分類番号のベクトル空間に射影された分散表現と、分類番号の分散表現との cos 類似度を求める必要があることから、語義の選択はその単語の文章中に存在した分類番号の中からのみ行われる問題がある。従って『分類語彙表』の辞書を用いて、コーパス中だけではなく、その単語がとりうる全ての分類番号を考慮する必要がある。しかし、新しく考慮する分類番号の分散表現が作成されているためには、その分類番号を持つ単語が一つでもコーパス中に存在する必要があり、全ての単語がその分類番号を持たない場合、その分類番号は単語の語義の選択に用いられない問題は残る。

本研究では、文章中に出現回数が 50 回以上出ている単語を対象に実験を行ったが、語義ごとの出現回数を数えると 1 回しか出現しない語義が 4 語義存在した。分散表現を作成するには、ある程度の用例数が必要であると考えられる。そのため、コーパス中の用例数による、語義の対応付けの性能がどのように異なるかを調べた。ここで、25 語義を 5 語義ずつ出現回数の

少ない順にまとめ、縦軸に正解/不正解数をとり、各群において正解数と不正解数を図 1 に出力した。図 1 の棒グラフのラベルは「“各群の最少の出現回数”-“各群の最多の出現回数”」を示す。

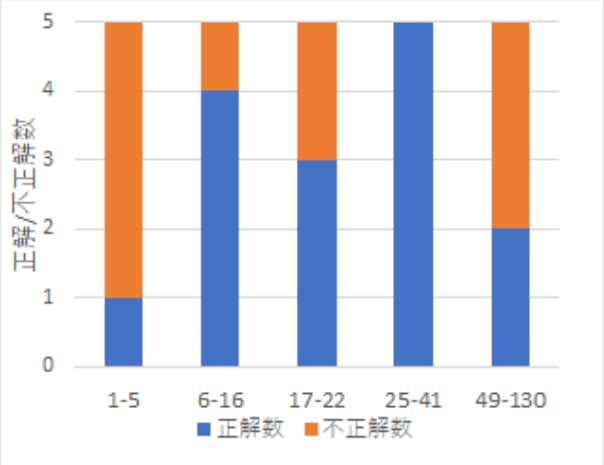

図 1: 出現回数毎の正解数

図 1 より、本研究による語義の推定では、出現回数が極端に少ない語義の正解率は低いが、必ずしも出現回数が多くなるほどより正確になるわけではないことがわかる。

今回使用したコーパス『現代日本語書き言葉コーパス』は分類番号が人手で付与されており、その精度は非常に高い。しかし、二種類以上のタグが付与されているコーパスはまだ少ないため、タグが一種類のみ付いているコーパスに、別のタグセットのタグをタガーによる付与を行い、提案手法を用いて対応をとり、その正解率の調査を進めていくことを今後考えている。

5 おわりに

本研究では、コーパス『現代日本語書き言葉コーパス』から、同一の文章に語義タグを付与し作成した分かち書き文から作成した分散表現と、分類番号を付与し作成した分かち書き文から作成した分散表現の間に BWE を適用することで『岩波国語辞典』と『分類語彙表』の対応をとる実験を行った。

その結果、提案手法では、最多出現語義を付与した場合よりも正解率が下がった。しかし、語義が分散表現で求まることから、分類番号を構成する語義タグの割合を調べられるという長所がある。

しかし、本研究の正解率は学習コーパスとテストコーパスが同じ範囲であることと、学習段階で学習した単語と語義タグと分類番号の組み合わせのみを正解としていることから評価対象を拡張させる必要がある。また、コーパスを拡張した際に本研究の正解率がどのように変化するか、追加調査していきたい。

謝辞

本研究は、JSPS 科研費 18K11421 の助成を受けたものである。

参考文献

[1] Karl Moritz Hermann and Phil Blunsom. Multilingual models for compositional distributed semantics. *arXiv preprint arXiv:1404.4641*, 2014.

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

[3] Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den. Balanced corpus of contemporary written japanese. *Language resources and evaluation*, Vol. 48, No. 2, pp. 345–371, 2014.

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

[5] 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.

[6] 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.