Bilingual Word Embeddingsによる 短単位と長単位のアラインメント

平林 照雄　　古宮 嘉那子　　新納 浩幸
茨城大学 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

一般に、文章を構成する言語単位は複数存在する。特に、日本語は単語境界が曖昧となることが多く、比較を行う対象語同士の単語数が異なることがしばしば起こりうる。例えば、国立国語研究所が提唱する短単位を用いると、「会員」は 1 単語であるが、「裁判員」は「裁判／員」と、2 単語からなり、これらを直接評価することが難しい。そこで、本研究では、通常、他言語間で同一の空間上に分散表現を作成する用途に用いられる Bilingual Word Embeddings(以下 BWE と略す) を、短単位に分割された文章から作成された分散表現と、長単位に分割された文章から作成された分散表現間に適応することで、異なる分散表現空間上の単語の分散表現の比較を行う。

2 関連研究

BWE 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。Mikolov ら [9] 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。Xiao と Guo ら [11] 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。Hermann と Blunsom ら [5] 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 [6] は、joint optimization の手法を初めて行った。

また本研究では、形態素情報を使用せずに実験を行ったが、Yang ら [12] は、形態素情報に注目したアラインメントを行うことで、教師なしでの訳語ペア作成ながら、教師ありに匹敵するほどの精度を得ている。

また、異なる単語数の単語同士を比較する手法として、複数の単語を 1 単語に合成し比較する手法が一般的である。Komiya ら [7] は単語ベクトルから句ベクトルを作成するモデルを利用して、13 の係り受け関係を設定し、各係り受け関係ごとにモデルを生成し、短単位から長単位の分散表現を生成している。

本研究は、単一言語マッピングの手法により BWE を構築し、分散表現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 を分散表現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、BWE を用いて短単位分散表現と長単位分散表現の対応を取る。短単位分散表現には nwjc2vec[10] を用い、長単位分散表現には『現代日本語書き言葉均衡コーパス』[8] の長単位コーパス (以下 BCCWJ と略す) の分かち書き文から word2vec [1] を使って生成した分散表現を用いる。BWE の構築には、正規化と線形変換のみを行う単純線形変換と、BWE フレームワークを構築するオープンソースである VecMap [2] を使用した。

[1] https://code.google.com/archive/p/word2vec/
[2] https://github.com/artetxem/vecmap#publications

3.1 nwjc2vec

nwjc2vec は新納ら [10] が『国語研日本語ウェブコーパス』（以下 NWJC と略す）[4] から作成した短単位分散表現辞書である。NWJC の統計データを表 1 に示す。また今回使用した nwjc2vec は、形態論情報を

表 1: NWJC の統計データ

収集 URL 数	83,992,556
文数 (のべ数)	3,885,889,575
文数 (異なり数)	1,463,142,929
国語研短単位数	25,836,947,421

含む分散表現が作成されていたが、BCCWJ では形態論情報を考慮せずに作成したため、nwjc2vec の形態論情報部分を削除した分散表現を BWE の構築に用いた。そのため、短単位分散表現内には同じ表層形で異なる品詞の分散表現が、同一のエントリで異なるベクトルを持つ分散表現として発生したが、それらの分散表現をそのまま保持して構築した。

3.2 BCCWJ

『現代日本語書き言葉均衡コーパス』[8] は、国立国語研究所によって作成された、複数のジャンルの文書が含まれた均衡コーパスである。それぞれの文書は短単位と長単位の 2 種類の単位により分割されている。コーパスの統計データを表 2 に示す。

表 2: BCCWJ の統計データ

サンプル数	172,675
短単位数	104,911,464
長単位数	83,585,665

3.3 VecMap

Artetxe ら [1][2][3] が作成した、BWE のスクリプト及び、単語の翻訳や類似性、関連性／類推を評価するツールを内包するオープンソースを VecMap のツールとして用いた。本実験ではこれらのうち、supervised オプションと identical オプションによる BWE を用いた各分散表現を共通分散表現空間上にマッピングを行った分散表現を作成する機能を用いる。

4 実験

4.1 評価実験

作成した分散表現の評価のため、『分類語彙表』[3] の分類番号を用いて、作成した分散表現の妥当性の検証を行った。『分類語彙表』は、語義が階層構造（木構造）の中で定義されている概念辞書であるため、同じノードに属する語義同士は距離が近くなることが予想される。これを利用し、作成した分散表現の評価を行った。また、以下説明のために、長単位区切りでは 1 単語であるが、短単位区切りでは 2 単語であるような語を「長単位語」、長単位区切りでは存在しないが、短単位区切りでは 1 単語であるような語を「短単位語」と呼称する。評価は以下の手順で行う。

1. 分類番号を持つ長単位語のうち、構成する短単位語 (ws_{i_1}, ws_{i_2}) も分類番号を持つ語 wl_i に対して、wl_i と同じ分類番号を持つ語群ノード $N_i(0)$ を定義する。

2. $N_i(0)$ 中の wl_i 以外の全ての短単位語の分散表現と wl_i の分散表現 e_i の cos 類似度の平均 $s_i(0)$ を求める。この時、分散表現は本研究で作成した分散表現を用いる。

3. 同様に分類番号語群ノード $N_i(n)$ に対し、ノード中の全ての短単位語の分散表現と e_i の cos 類似度の平均 $s_i(n)$ を求める。

4. $s_i(n)$ の値を大きい順に順位付けした時の $s_i(0)$ の順位求めた。

また比較実験として、e_i を、(ws_{i_1}, ws_{i_2}) の平均分散表現とし、$N_i(n)$ 中の短単位語の分散表現に nwjc2vec の分散表現を用いる実験（以下、「平均」と記す）と、e_i を、wl_i を構成する短単位語のうち、後ろの語 ws_{i_2} の分散表現とし、$N_i(n)$ 中の短単位語の分散表現に nwjc2vec の分散表現を用いる実験（以下、「後ろの語」と記す）を行った。

4.2 実験設定

本実験で用いた BCCWJ と NWJC の統計データを表 3 に、word2vec のパラメータを表 4 に示す。

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

表 3: BCCWJ と NWJC の統計データ

BCCWJ 長単位単語種類数	2,745,657
nwjc2vec 短単位単語種類数	1,534,957
共通語	278,143

表 4: word2vec のパラメータ

次元数	200
学習アルゴリズム	C-BoW
ウィンドウ幅	5
反復回数	5
バッチサイズ	1,000
min-count	1

単純線形変換に用いた線形変換 W の学習パラメータを表 5 に示す。

表 5: W のパラメータ

次元数	200 × 200
損失関数最適化アルゴリズム	Adam
反復回数	1164

W のパラメータの反復回数は、訓練データから 55,630 語をランダムに評価用データとしてとり、最小の loss となる反復回数を調査した。また、この調査を 5 回行い、その平均反復回数とした。

また、単純線形変換による BWE の構築に用いた分散表現数及び、そのシード単語数を表 6 に示す。

表 6: 単純線形変換による BWE の構築に用いる分散表現種類数

BCCWJ 長単位単語種類数	2,745,657
nwjc2vec 短単位単語種類数	1,534,957
シード単語種類数	278,143

VecMap で使用した分散表現数及び、supervised オプションで用いた seed dictionary の数を表 7 に示す。

表 7: VecMap による BWE の構築に用いる分散表現種類数

BCCWJ 長単位単語種類数	289,805
nwjc2vec 短単位単語種類数	1,534,957
seed dictionary	278,143

VecMap で使用した長単位分散表現には、共通語 278,143 語に、評価実験に用いる専門家の人手によって分類番号を付与した 11,662 語を加えた 289,805 語を用いた。

評価実験に用いた wl_i の語数 i は専門家の人手によって分類番号を付与した 11,662 語のうち、$N_i(0)$ 中の語すべてに分散表現を生成されていない 203 語を除いた 11,459 語となる。また、分類番号語群ノード数は、ノード中の語のすべてに分散表現が生成されていない 14 を除いた 881 となる。

4.3 実験結果

実験の対象の 11,459 語に対して評価実験を行った時の $s_i(0)$ の平均順位を、提案手法（単純線形変換、supervised、identical)、実験（平均)、実験（後ろの語）の順に表 8 に示した。

表 8: 各手法における正解分類番号ノードの平均順位

手法	平均順位
提案手法 (単純線形変換)	187.50 位
提案手法 (supervised)	131.98 位
提案手法 (identical)	330.40 位
実験 (平均)	80.41 位
実験 (後ろの語)	143.16 位

5 考察

表 8 と、分散表現がランダムで与えられたときの平均順位は全体順位の半分の 440.5 位となることから、BWE による異なる分散表現の対応の正確さはランダムより高いが、最も正確だった vecmap の supervised オプションによる対応でも、実験 (平均) を上回ることは出来ず、一つの分散表現辞書から、未知語の分散表現を、構成する単語の分散表現の平均とした時の手法が最も正確であることがわかった。

6 おわりに

本研究では、BWE を用いて短単位分散表現と長単位分散表現を同一分散表現空間上にマッピングを行った。また、『分類語彙表』の木構造を利用した評価を行った。その結果、既存の手法を超えることができず、

未知語の分散表現を、構成する単語の分散表現の平均とした時が最も正確であることが分かった。

謝辞

本研究は、茨城大学の若手教員研究費支援制度「分散表現を用いた多単語表現の変換」および JSPS 科研費 18K11421 の助成を受けたものである。

参考文献

[1] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 451–462, 2017.

[2] 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*, pp. 5012–5019, 2018.

[3] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 789–798, 2018.

[4] Masayuki Asahara, Kikuo Maekawa, Mizuho Imada, Sachi Kato, and Hikari Konishi. Archiving and analysing techniques of the ultra-large-scale web-based corpus project of ninjal, japan. *Alexandria*, Vol. 25, No. 1-2, pp. 129–148, 2014.

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

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

[7] Kanako Komiya, Takumi Seitou, Minoru Sasaki, and Hiroyuki Shinnou. Composing word vectors for japanese compound words using dependency relations. *CICLING*, 2019. no 229.

[8] 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.

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

[10] 新納浩幸, 浅原正幸, 古宮嘉那子, 佐々木稔. nwjc2vec: 国語研日本語ウェブコーパスから構築した単語の分散表現データ. 自然言語処理, Vol. 24, No. 5, pp. 705–720, 2017.

[11] 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.

[12] Pengcheng Yang, Fuli Luo, Peng Chen, Tianyu Liu, and Xu Sun. MAAM: A morphology-aware alignment model for unsupervised bilingual lexicon induction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3190–3196, Florence, Italy, July 2019. Association for Computational Linguistics.