# 教師なし機械翻訳の出力結果から作成した Bilingual Word Embeddingsの分析

西川 荘介　　鶴岡 慶雅
東京大学 工学部電子情報工学科
{sosuke,tsuruoka}@logos.t.u-tokyo.ac.jp

## 1 はじめに

単語の意味をベクトル化して表す、単語埋め込み表現 (Word Embedding) [10] が自然言語処理の様々な分野に応用されている。その中でも複数言語の単語埋め込み表現を共通の意味空間上で扱う Bilingual Word Embeddings (BWE) を利用することで、英語などの資源の豊富な言語で訓練したモデルを他の言語に転用できることが報告されている。

BWE の応用例の 1 つに教師なし機械翻訳がある。2018 年に発表された Lample らの教師なし機械翻訳モデル [7] では、ドメインは同じであるが対訳関係のない多言語コーパス（コンパラブルコーパス) から作成した BWE を利用することである程度の精度を持つ機械翻訳モデルの作成に成功している。

上述のように教師なしで BWE を作成するにはコンパラブルコーパスが必要であるが、一般に提供されているコンパラブルコーパスの量には限りがある。本研究では教師なし機械翻訳機の出力結果を再利用することで元の訓練データを拡張し BWE を作成する手法を提案する。

提案手法にて作成した BWE はいくつかの言語横断タスクにおいて既存手法を上回る結果を示した。

## 2 関連研究

### 2.1 Bilingual Word Embeddings

単語埋め込み表現 [10] は似た意味の単語は似た文脈で出現するという分布仮説 [16] に基づき、単語の共起情報を学習することで取得されるが、異なる言語において似た意味を表していたとしても別々に訓練されたベクトル空間でのベクトルの類似度が高くなるとは限らない。

しかし言語が異なっていても単語埋め込み空間内の幾何学的関係が言語間で類似しているという前提の下、似た意味を表す単語は類似度が高くなるように適切な線形変換を用いて共通のベクトル空間上に写像した表現を Bilingual Word Embeddings (BWE) と呼ぶ [11]。Artetxe ら [2] は単語の類似度分布と自己学習の利用によりコンパラブルコーパスのみを利用して単語辞書なしで教師あり学習に匹敵する精度を持つ BWE の生成に成功した。本研究では拡張した訓練データを用いて作成した BWE の分析を行う。

### 2.2 教師なしフレーズベース機械翻訳

Lample [7] らにより提案された教師なしフレーズベース機械翻訳では, 訓練コーパスとして一切の対訳データを用いず, コンパラブルコーパスのみを利用して翻訳システムを実現した. 以下に学習アルゴリズムの概要を示す。

#### フレーズテーブルの作成

コンパラブルコーパスから BWE を作成し、BWE を利用して以下の式に従ってソース文のあるフレーズから翻訳される可能性の高いターゲット文のフレーズを記した表 (フレーズテーブル) を作成する。

$$p(t_j|s_i) = \frac{\exp\left(\frac{1}{T}\cos(e(t_j), e(s_i))\right)}{\sum_k \exp\left(\frac{1}{T}\cos(e(t_k), e(s_i))\right)} \quad (1)$$

ここで $t_j$ はターゲット言語の $j$ 番目のフレーズを表し、$s_i$ はソース言語の $i$ 番目のフレーズを表すまた、$T$ は分布のピーク調整のためのハイパーパラメータを表し、$e(x)$ は $x$ の BWE を表す[1]。

#### 言語モデルの作成

n-gram 言語モデルである KenLM [17] を利用して言語モデルを作成する。

#### 逆翻訳

作成したフレーズテーブルでフレーズごとに翻訳し、言語モデルを利用してそれらが自然に並び替えられるようにすることで最初のソース言語からターゲット言語への翻訳モデル $P^0_{s\to t}$ を作成し、そのモデルの生成結果と元の訓練コーパスを擬似対訳コーパスとして教師あり学習 [13] によりターゲット言語からソース言語への翻訳モデル $P^1_{t\to s}$ を生成する。同様にして翻訳モ

[1] 2 単語以上を用いる場合はフレーズの埋め込み表現 [3] を利用する。

デルからの擬似対訳コーパスの生成と、それを用いた教師あり学習による翻訳モデルの生成を繰り返す逆翻訳 [15] を行うことで翻訳モデルの精度を向上させる。

本研究ではこの教師なしフレーズテーブル機械翻訳機を用いることで単語埋め込み表現学習用の訓練データを拡張する。

### 2.3 教師なし機械翻訳の生成結果の利用

教師なし機械翻訳の結果を利用する研究はいくつか他にも行われている。Marie ら [9] は元の訓練データと教師なしフレーズベース機械翻訳結果から得られる擬似対訳コーパスに対して、bilingual skipgram [8] を用いて学習した BWE の有効性を示している。また Artetxe ら [1] は Marie らと同様にして取得した擬似対訳コーパスに対して FastAlign [5] にて単語の対応関係を学習させ単語辞書を取得している。本研究ではこれらの研究とは異なり、教師なし機械翻訳の結果を対訳コーパスではなく元の訓練データの拡張データとして連結し、単語埋め込み表現の学習と写像により再度 BWE を学習する手法を提案する。

## 3 提案手法

以下に本提案モデルによる BWE 獲得手法を示す。

#### 教師なし機械翻訳機の学習

コンパラブルコーパスの各言語のデータ（以下訓練コーパスと呼ぶ) を用いて、512 次元の単語埋め込み表現を fastText[2]を用いて獲得し、Artetxe らの教師なしの手法 [2] により同一空間上に写像することで BWE を獲得する。BWE を利用することで Lample らの手法 [7] により教師なしフレーズテーブル機械翻訳機[3]を作成した後、逆翻訳を繰り返すことでソース言語とターゲット言語両方向の教師なし機械翻訳機を訓練する。

#### Bilingual Word Embedding の再獲得

作成した教師なし機械翻訳機へ訓練コーパスを入力し、その出力結果（以下擬似コーパスと呼ぶ) を獲得する。各言語、擬似コーパスを元の訓練コーパスと連結し、再度同じ手順で BWE を学習する。Lample ら [7] の設定では教師データとして単語辞書を用いて BWE を作成しているが本実験では全ての訓練過程で教師データを一切用いていない。

[2]https://github.com/facebookresearch/fastText

[3](1) 式において T=30 に設定し、ソース言語における頻出語 30000 語それぞれに対して (1) 式の値の高い 200 語を抽出することでフレーズテーブルを作成した。

表 1: 教師なし機械翻訳機の BLEU スコア

| $en \to fr$ | $fr \to en$ | $en \to de$ | $de \to en$ |
|---|---|---|---|
| 19.28 | 19.03 | 10.34 | 13.73 |

## 4 実験・結果

本研究では BLI、言語横断タスク（文書分析･感情分析)、意味的類似度タスクにて提案手法の評価を行った.

訓練コーパスは Wikipedia Comparable Corpora[4] から英語、フランス語、ドイツ語を 1000 万文ずつ利用した。使用した全てのコーパス、データセットに対して、mosesdecoder[5]により単語分割を行い小文字化した。

擬似コーパスの質がどの程度かを確かめるために擬似コーパスの生成に利用した教師なし機械翻訳機の BLEU スコア [12] を算出した。BLUE スコアの結果を表 1 に示す。BLUE スコアの算出には WMT14[6] より newstest2014 における英仏、英独対訳コーパスを用いている。

全ての実験において拡張前と拡張後の単語埋め込み表現において共通する語彙 50000 単語を抽出している。また、BLI の実験以外での提案手法は全てソース言語、ターゲット言語ともに拡張している。

#### Bilingual Lexicon Induction

Bilingual Lexicon Induction (BLI) とは BWE において似た意味を持つ各言語の単語がどれだけ近い位置にあるかを評価するタスクの 1 つである。人手による翻訳がなされた単語辞書データをテストデータとして用い、各テストソース単語に対して CSLS[7]が高い順にターゲット単語を順位付けし、対応するテストターゲット単語の順位の逆数の平均を評価値として用いる。

以下のコーパスにおいてそれぞれ単語埋め込み表現を学習し、全てのソース・ターゲットペアに対して 3 回異なる乱数シードで BWE を学習し BLI での性能の平均と標準偏差を算出した。

翻訳された文章は翻訳元の言語の性質を反映することがある。そこで本実験ではソース言語の写像先であるターゲット言語の文法構造・特徴的な表現をある程度保持した擬似コーパスによる拡張だからこそ、単語埋め込み表現の学習の際に単語の共起情報がより似通うことで幾何学的構造がより近くなり写像精度が向上するのではないかと考え、これを検証するために写像先ではない言語からソース文を拡張する実験も行った。

[4]https://linguatools.org/tools/corpora/wikipedia-comparable-corpora/

[5]https://github.com/moses-smt/mosesdecoder

[6]http://www.statmt.org/wmt14/translation-task.html

[7]高次元における hubness 問題 [14] の解決策として Conneau ら [4] に考案された類似度の指標

表 2: BLI の結果

| コーパス | fr | fr + psfr | de | de + psde |
|---|---|---|---|---|
| en | 0.655 ± 0.003 | 0.685 ± 0.001 | 0.548 ± 0.004 | **0.615**±0.002 |
| en + psen(fr) | **0.704**±0.001 | **0.693**±0.000 | 0.561 ± 0.005 | 0.614 ± 0.004 |
| en + psen(de) | 0.647 ± 0.004 | 0.680 ± 0.002 | **0.610**±0.004 | 0.611 ± 0.004 |

表中の psen(fr) はフランス語によって生成された英語の擬似コーパスを表す。

- ソース言語（英語）訓練コーパスのみ
- ソース言語（英語）訓練コーパス
  ＋（フランス語 or ドイツ語）から生成した擬似コーパス
- ターゲット言語（フランス語 or ドイツ語）
  訓練コーパスのみ
- ターゲット言語（フランス語 or ドイツ語）
  訓練コーパス ＋ 英語から生成した擬似コーパス

テストデータセットは xling-eval [6] を用いており、英語・フランス語のテストペア数は 1881、英語・ドイツ語のテストペア数は 1686 で BLI の精度を計算している。

BLI の結果を表 2 に示す。結果から基本的には擬似コーパスによる拡張を行うことで BLI の精度が上昇することが確認された。また、ソース訓練コーパスを BWE の写像先ではない他の言語から生成された擬似コーパスによって拡張した場合では BLI の精度が減少する場合もあることが確認された。この結果は本手法が単なるデータ拡張ではなく写像先の言語による影響があるという仮説を支持する。

### 言語横断タスク

本研究では言語横断タスクとして文書分類タスクと感情分析タスクにて BWE の評価を行った。両タスクでは英語を訓練コーパスとし、フランス語、ドイツ語をテスト、開発コーパスとして 10 回乱数シードを変えて実験を行い、精度の平均と標準偏差を算出し、有意水準 1%の t 検定を行った。

文書分類タスクは文章をトピックに応じて分類するタスクである。 データセットは MLDoc[8]を用いた。各文書は CCAT (Corporate/Industrial)、ECAT (Economics)、GCAT (Government/Social)、MCAT (Markets) のいずれかのカテゴリが付与されている。データセットの詳細を表 3 に示す。

感情分析タスクはレビュー文章などにおいて文章が肯定的な意見か否定的な意見かを分類するタスクである。データセットとして amazon のレビュー[9]を用いた。このデータセットは amazon の商品に対する 1 から 5 までの評価値データとレビュー文章から成り、評価値 1-2 を “negative”、評価値 4-5 を “positive” とし、評価値が 3 のレビュー文書は除外した。データセットの詳細を表 4 に示す。

[8]https://github.com/facebookresearch/MLDoc
[9]https://webis.de/data/webis-cls-10.html

表 3: 文書分類に用いたデータセット

| カテゴリ | 訓練 | テスト (fr/de) | 開発 (fr/de) |
|---|---|---|---|
| CCAT | 8761 | 3996/3936 | 1218/1276 |
| ECAT | 8785 | 3892/4104 | 1251/1253 |
| GCAT | 8799 | 3992/4088 | 1280/1214 |
| MCAT | 8656 | 4120/3872 | 1249/1256 |

表 4: 感情分析に用いたデータセット

| 感情 | 訓練 | テスト | 開発 |
|---|---|---|---|
| negative | 2625 | 3000 | 375 |
| positive | 2625 | 3000 | 375 |

表 5: 文書分類タスクの結果

| コーパス | en-fr | en-de |
|---|---|---|
| 擬似コーパスなし | 78.8 ± 1.5 | 78.5 ± 1.6 |
| 擬似コーパスあり | **82.8**$^{\dagger}$ ± 1.6 | 78.8 ± 2.7 |

表 6: 感情分析タスクの結果

| コーパス | en-fr | en-de |
|---|---|---|
| 擬似コーパスなし | 69.3 ± 0.6 | 63.4 ± 1.3 |
| 擬似コーパスあり | 68.6 ± 1.9 | **64.9**$^{\dagger}$ ± 1.0 |

†既存手法と有意差があること（$p < 0.01$）を表す。

文章分類タスク、感情分析タスクの結果を表 5、6 に示す。文章分類タスクでは英語・フランス語間にて提案手法が既存手法のスコアを上回ることが確認されたが英語・ドイツ語間では有意差は認められなかった。感情分析タスクでは英語・フランス語間では有意差は認められなかったが英語・ドイツ語間では既存手法のスコアを上回ることが確認された。これらの結果からは一貫した傾向が得られておらず、他言語でも追試を行う必要がある。

**意味的類似度タスク (Word Similarity)**

意味的類似度タスクは人手により作成された単語ペアの類似度テストデータと単語埋め込み表現から計算されるコサイン類似度の相関を計測することで単語埋め込み表現としての質を評価するタスクである。英語にて 6 回異なる乱数シードで作成した既存手法と提案手法による単語埋め込み表現を用いて実験を行い、相関の平均と標準偏差を算出し、有意水準 1%の t 検定を行った。

データセットとして動詞 3500 ペアで構成された simverb-3500[10] と web 上でクローリングしたテキストから抽出された頻出語 3000 ペアで構成された men[11] を用いた。

意味的類似度タスクの結果を表 7 に、結果の例を表 8 に示す。simverb-3500 では有意差は認められなかったが、men ではいずれの言語においても提案手法のスコアが既存手法を若干上回った。この結果から本手法は BLI や言語横断タスクへの応用だけでなく、単言語埋め込み表現自体の質を向上させる可能性を示している。

表 7: 意味的類似度タスクの結果

| コーパス | simverb-3500 | men |
|---|---|---|
| 擬似コーパスなし | $0.259 \pm 0.006$ | $0.763 \pm 0.001$ |
| 擬似コーパスあり (fr) | $0.260 \pm 0.004$ | $\mathbf{0.767}^{\dagger} \pm 0.002$ |
| 擬似コーパスあり (de) | $0.253 \pm 0.003$ | $\mathbf{0.768}^{\dagger} \pm 0.002$ |

†既存手法と有意差があること（$p < 0.01$）を表す。

表 8: 意味的類似度タスク（men）の結果の例

| word1 | word2 | sim | cossim |
|---|---|---|---|
| sun | sunlight | 50 | 0.472 |
| guitar | music | 40 | 0.462 |
| cold | washing | 20 | 0.162 |
| chair | ipod | 5 | -0.019 |

sim は人手による類似度、cossim はコサイン類似度を表す。

## 5 おわりに

本研究では教師なし機械翻訳の出力結果を利用することで訓練データを拡張し BWE を作成する手法を提案した。実験ではこの拡張手法による BWE が BLI、一部の言語横断タスク、一部の意味的類似度タスクにおいて既存手法の性能を上回ることを示した。

今後の展望としては提案手法について他の種類の言語横断タスクでの性能検証、教師なし機械翻訳への再利用の検討が考えられる。

[10] http://people.ds.cam.ac.uk/dsg40/simverb.html
[11] https://staff.fnwi.uva.nl/e.bruni/MEN

## 参考文献


[1] Artetxe et al. Bilingual lexicon induction through unsupervised machine translation. *In ACL, 2019.*

[2] Artetxe et al. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. *In ACL, 2018.*

[3] Cho et al. Learning phrase representations using RNN encoder–decoder for statistical machine translation. *In EMNLP, 2014.*

[4] Conneau et al. Word translation without parallel data. *In ICLR, 2018.*

[5] Dyer et al. A simple, fast, and effective reparameterization of IBM model 2. *In ACL, 2012.*

[6] Glava et al. How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. *In ACL, 2019.*

[7] Lample et al. Phrase-based & neural unsupervised machine translation. *In EMNLP, 2018.*

[8] Luong et al. Bilingual word representations with monolingual quality in mind. *In VSMNLP, 2015.*

[9] Marie et al. Unsupervised joint training of bilingual word embeddings. *In ACL, 2019.*

[10] Mikolov et al. Efficient estimation of word representations in vector space. *In ICLR, 2013.*

[11] Mikolov et al. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*, 2013.

[12] Papineni et al. Bleu: a method for automatic evaluation of machine translation. *In ACL, 2002.*

[13] Philipp et al. Statistical phrase-based translation. *In ACL, 2003.*

[14] Radovanović et al. Hubs in space: Popular nearestneighbors in high-dimensional data. *Journal of Machine Learning Research,11,2010.*

[15] Sennrich et al. Improving neural machine translation models with monolingual data. *In ACL, 2016.*

[16] Zellig S. Harris. Distributional structure, word, 10. *1954.*

[17] Kenneth Heafield. KenLM: Faster and smaller language model queries. *In WMT, 2011.*