卒業論文

教師なし機械翻訳の出力結果から作成した多言語埋め込み表現の学習

令和 2 年 2 月 7 日提出

指導教員　鶴岡 慶雅 教授

電子情報工学科

03180432 西川 荘介

目次

概要

近年、自然言語処理の様々なタスクでは英語を中心とする資源が豊富な言語にて大幅な精度の向上が報告されている一方、学習に十分な資源が確保できない言語での自然言語処理タスクの精度は比較良くないものが多い。

以上の問題に対して、異なる言語の単語を共通意味空間上の数値ベクトルとして扱うことができる多言語埋め込み表現を利用することで資源の豊富な言語にて訓練したモデルをその他の言語に転用できることが報告されている。

多言語埋め込み表現の学習では独立に学習された各言語の単語埋め込み表現を線形変換を用いて共通のベクトル空間上に写像するが、線形写像のみでの学習は難しく、利用できる訓練コーパスの量には限りがある。

そこで本研究では教師なし機械翻訳結果を元の訓練コーパスの拡張データとして用いることで多言語埋め込み表現を学習する手法を提案する。翻訳元の性質を反映した文章による拡張により、より各言語の単語埋め込み表現学習時の単語共起情報が似通い、単語埋め込み空間の幾何学的構造が近づくことで、写像性能が向上することを期待している。

本研究では提案手法による多言語埋め込み表現の学習において写像性能が向上していることを示し、さらに写像先ではない言語による拡張との精度の比較により、上記の仮説を検証した。また、提案手法による多言語埋め込み表現は既存手法と比べ言語横断タスクの性能を向上させる可能性や単言語埋め込み表現の質を維持していることを示し、教師なし機械翻訳への再利用の可能性についても検討した。

第 1 章

はじめに

1.1 背景

近年、自然言語処理では英語をはじめとする言語資源が豊富な言語において膨大なデータを用いた機械学習による研究が進み、様々なタスクでの性能が劇的に向上している。一方で日本語や韓国語など、英語を含む欧米系の言語から比較的関係性が遠い言語や、利用人数の少ない言語においては研究があまり進んでおらず実際のタスクの性能もあまり良くないものが多い。

上記の問題に対して、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) が自然言語処理の様々な分野に応用されているが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多言語埋め込み表現を利用することで、英語などの資源の豊富な言語にて訓練したモデルをその他の言語に転用できることが報告されている [1]。

多言語埋め込み表現の学習では各言語の単語埋め込み表現を独立して学習したのちに、単語辞書を利用して学習した変換行列を用いてソース言語の単語埋め込み空間からターゲット言語の単語埋め込み空間へ線形写像を行うことで共通のベクトル空間で扱う手法 [2] が提案されている。近年では単語辞書を利用せずに教師なしで写像を学習する手法も考案されている [3, 4]。これらの手法は言語が異なっていても、単語埋め込み空間の幾何学的構造が似ている前提の下、成り立つ手法であるが、異なる言語の単語埋め込み空間の幾何学的構造が一般的に必ずしも似通うわけではないことが報告されている [5]。また、教師なしで単語の写像を行うに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) が必要とな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り、良い多言語埋め込み表現の学習には限界がある。

多言語埋め込み表現の応用例の 1 つとして教師なし機械翻訳が考えられる。本来、機械翻訳が十分な性能を発揮するには人手により翻訳されたデータである対訳コーパスが大量に必要であるが、言語によっては対訳コーパスが十分に存在しないこともあり、新たな対訳コーパスの作成には多大なコスト、時間がかかる。以上の問題を解決するため、2018 年に発表された Lample らの教師なし機械翻訳モデル [6] では、多言語埋め込み表現を利用することで対訳コーパスを一切用いず、コンパラブルコーパスのみで一定の精度を持つ翻訳モデルの作成に成功している。

1.2 本研究の貢献

本研究の貢献は以下の通りである。

- 教師なし機械翻訳モデルの出力結果である擬似コーパスにて、限られた量のコンパラブルコーパスを拡張することで作成した多言語埋め込み表現において、単語の写像精度や一部の言語横断タスクの精度が向上することを示した。
- 翻訳元の言語の性質を保持した擬似コーパスによる単語埋め込み表現の学習は、各言語の単語埋め込み空間の幾何学的構造を近づけることで、多言語埋め込み表現の学習精度を向上させる可能性を示した。
- 擬似コーパスを利用して獲得した多言語埋め込み表現から作成される翻訳モデルは出力結果の多様性が

低く、逆翻訳による翻訳モデルの精度向上率が下がることを示した。

第 2 章

関連研究

2.1 単語埋め込み表現

自然言語処理においては単語をコンピュータで計算するために単語の特徴を数値ベクトル化して表すことが多い。最もシンプルな表現方法としてベクトルの各要素を文章全体に現れる全ての単語に対応させて、各単語のベクトルを自分自身の単語の要素は 1、それ以外の要素は 0 とする One-Hot 表現が考えられる。例えば “I like a crane very much”という文書について、“like”の One-Hot 表現は以下のように表せる。

$$\boldsymbol{x_{like}} = (0, 1, 0, 0, 0, 0) \tag{2.1}$$

しかし One-Hot 表現ではベクトルの次元が文書の語彙数に依存するので文書によっては次元が巨大になり、ベクトルが疎（要素のほとんどが 0) になってしまう。このようなベクトルを利用する場合必要な統計値が十分に獲得できないことがある。またこのベクトルは単語の意味を捉えることができておらず、単語間の相対的な意味関係を表現できない。

近年の自然言語処理における深層学習では単語を連続値をとる低次元のベクトルとして扱うことでこの問題を解決している。これは単語埋め込み表現 (Word Embedding) と呼ばれ、Harris らの「単語の意味はその単語が使われた周囲の文脈によって決まる」という分布仮説 [7] に基づき、後述する skip-gram や CBoW [2] などのアルゴリズムにより単語の共起情報を学習することで獲得される。この単語埋め込み表現空間では似た意味の単語が近くなる。その例を図 2.1 に示す。また、図 2.2 に示すように以下の式 (2.2) のような単語ベクトル同士の演算も可能となる。

$$\boldsymbol{x_{king}} - \boldsymbol{x_{man}} + \boldsymbol{x_{woman}} = \boldsymbol{x_{queen}} \tag{2.2}$$

以下では単語埋め込み表現を獲得する手法について述べる。

skip-gram

skip-gram はある単語が与えられた際に、その周辺の単語を予測するモデルである。例えば “I like a crane very much”という文書が与えられた時、I の次にいきなり crane や very 等が現れる確率は低く、like やその他の動詞が出現しやすい。このようにどういった単語が次に出現しやすいかという確率を考えることができる。

上記のようにある注目単語 w_I の周辺に関連する単語 w_o が出てくる確率を以下のような条件付き確率の式で表せる。

$$p(w_o|w_I) = \frac{\exp\left(\boldsymbol{v}'^{T}_{w_o} \cdot \boldsymbol{v}_{w_I}\right)}{\sum_{w_v \in V} \exp\left(\boldsymbol{v}'^{T}_{w_v} \cdot \boldsymbol{v}_{w_I}\right)} \tag{2.3}$$

ここでは $\boldsymbol{v}$、$\boldsymbol{v}'$ は入力・出力ベクトル、$\boldsymbol{v}'^{T}$ は $\boldsymbol{v}$ の転置、V は文書の全語彙を表す。

さらに単語だけでなく、周辺複数単語を推測するモデルとして以下のようなコンテキスト C を導入することで図 2.3 のように周辺 C 個を予測する同時確率の式に拡張できる。この確率の値 p が最大になるような単語ベクトル v を単語埋め込み表現として取得する。

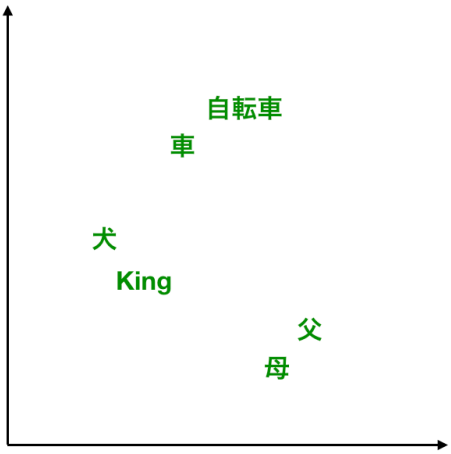

図 2.1. 単語埋め込み表現を二次元平面で表した概念図

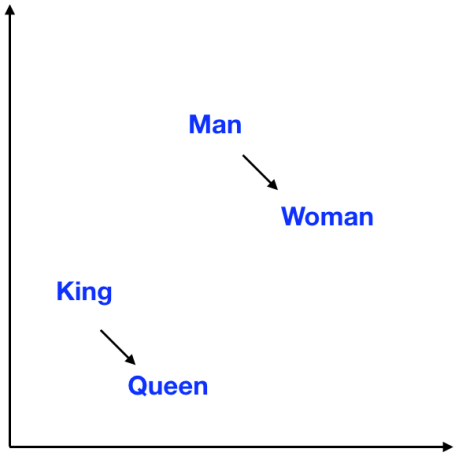

図 2.2. 単語埋め込み表現の相対関係を表した概念図

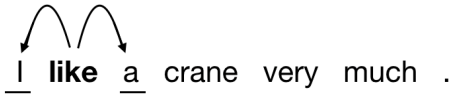

図 2.3. 近傍単語を推測する概念図

$$p(w_{o,1}, w_{o,2}, \cdots, w_{o,C}|w_I) = \prod_{c=1}^{C} \frac{\exp\left(\boldsymbol{v}'^{T}_{w_o} \cdot \boldsymbol{v}_{w_I}\right)}{\sum_{w_v \in V} \exp\left(\boldsymbol{v}'^{T}_{w_v} \cdot \boldsymbol{v}_{w_I}\right)} \tag{2.4}$$

すなわち、式 (2.5) のような目的関数をおき、確率的勾配降下法 (stochastic gradient descent, SGD) にて最適化を行う。

$$E = -\log p(w_{o,1}, w_{o,2}, \cdots, w_{o,C}|w_I) \tag{2.5}$$

ニューラルネットワークによる単語埋め込み表現の学習

単語埋め込み表現学習の実装として良く用いられる Word2Vec[*1]ではニューラルネットワークを用いて skip-gram が実装されている。本項ではその実装の概要を記述する。

図 2.4 のような全結合層のみからなる 2 層のニューラルネットワークを構築し、注目単語 w_i を表す i 要素目が 1 の One-Hot ベクトルを入力層に入力すると式 (2.6) に従って行列 W をかけることで隠れ層の出力 $\boldsymbol{v}_i$ が得られる。この $\boldsymbol{v}_i$ が注目単語 w_i を表す単語埋め込み表現となる。出力層にはコンテキストの数分、ユニット $\boldsymbol{u}$ が存在し、式 (2.7) に従って行列 W' をかけることで c 番目の $\boldsymbol{u}_c$ が得られる。出力層での活性化関数であるソフトマックス関数により式 (2.8) となりこれが skip-gram でモデル化した際の確率式 (2.3) と一致する。

$$\boldsymbol{h} = W\boldsymbol{x}_i = \boldsymbol{v}_i \tag{2.6}$$

$$\boldsymbol{u}_c = W'\boldsymbol{v}_i \tag{2.7}$$

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

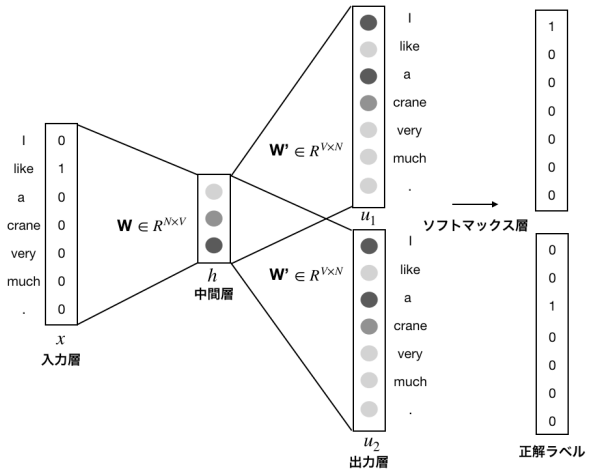

図 2.4. ニューラルネットワークによる単語埋め込み表現の学習

図中の V は文章の語彙数、N は隠れ層の次元を表し、W、W' はそれぞれ入力層、出力層への重みを表す。

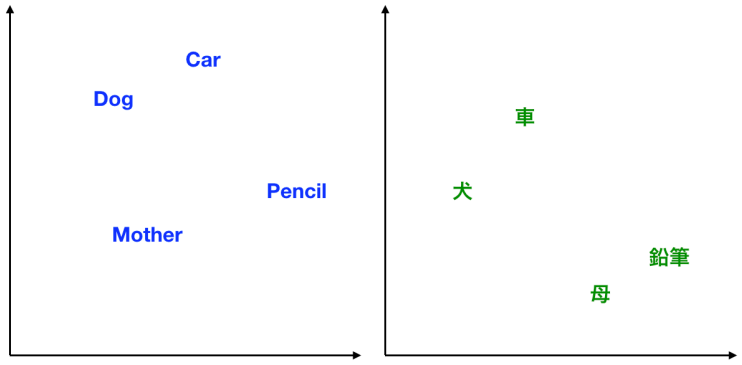

図 2.5. 異なる言語の単語埋め込み表現を二次元平面で表した際の概念図

$$y_{c,i} = \frac{\exp\left(\boldsymbol{u}_{c,i}\right)}{\sum_{v=1}^{V}\exp\left(\boldsymbol{u}_{c,v}\right)} = \frac{\exp\left(\boldsymbol{v}_i'^{T}\cdot\boldsymbol{v}_{w_I}\right)}{\sum_{v=1}^{V}\exp\left(\boldsymbol{v}_v'^{T}\cdot\boldsymbol{v}_{w_I}\right)} \tag{2.8}$$

その他の手法

Word2vec ではさらに負例サンプリング (Negative Sampling) と呼ばれる手法を導入し高速化を行っている。同じく単語埋め込み表現を学習する fastText [8] では単語から抽出した n-gram 全てに対して、上記の手法により単語埋め込み表現を割当て、n-gram の単語埋め込み表現の平均をその単語の埋め込み表現とする。例えば英単語 where に対して tri-gram で学習する場合を説明する。接頭辞、接尾辞であることを情報として持たせるために特殊な境界線記号<、>を用いて<where>とした後、<wh、whe、her、ere、re>と全体を表す<where>について同様に skip-gram アルゴリズムで各シンボルの単語埋め込み表現を学習した後、その平均を where の単語埋め込み表現とする。これによって新規単語や低頻出単語に対して適切なベクトル表現を得る可能性が高くなった。

2.2　多言語埋め込み表現

線形写像の学習

異なる言語においてそれぞれの単語埋め込み表現は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。そこで異なる言語において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多言語埋め込み表現 (Bilingual Word Embeddings) と呼ぶ。Mikolov ら [9] は図 2.5 に示すように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点に注目し、ソース言語の単語埋め込み空間からターゲット言語の単語埋め込み空間へ以下の式で表される平均二乗誤差を最適化することで学習される変換行列 W を用いた線形変換を行うことで多言語埋め込み表現を学習する手法を提案した。以下の式において x_i^s, x_i^t はテスト単語辞書にある単語ペアの i 番目のソース、ターゲット単語の単語埋め込み表現を表す。

$$\omega_{MSE} = \sum_{i=1}^{n} \|Wx_i^s - x_i^t\|^2 \tag{2.9}$$

Bilingual Lexicon Induction

Bilingual Lexicon Induction (BLI) は多言語埋め込み表現が各言語の単語埋め込み表現を共通空間上にどれだけ良く写像できているかを評価するタスクとして広く用いられている。

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の埋め込み表現に対してコサイン類似度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。例えば犬と dog がテストデータにあった場合、犬の単語ベクトル $\boldsymbol{x_{犬}}$ にコサイン類似度が高い順にターゲット単語ベクトルを順位付けした際、正解ターゲット単語ベクトル $\boldsymbol{y_{dog}}$ が全ターゲット単語の中で 3 番目にコサイン類似度が高かった場合、この逆数である $\frac{1}{3}$ を評価値とする。これにより正解ターゲット単語の順位が高い（コサイン類似度が高い) ほど評価値が大きくなる。これを全てのテストデータに対して行い、その平均を評価値として用いる。

変換行列の直行制約

式 (2.9) の学習では行列 W を直行行列とすることで単語埋め込み空間内の幾何学的構造が保たれ、写像精度が向上する。Søgaard ら [5] は W に直行制約を課さずに学習した多言語埋め込み表現は BLI の精度が高くても応用的な言語横断タスクにおいて性能を発揮しない場合があることを示した。

CSLS

Conneau ら [3] は多言語埋め込み表現の単語間の類似度として高次元における hubness 問題 [10] を解決するために Cross-Domain Similality Local Scaling (CSLS) と呼ばれる類似度の尺度を考案した。あるソース単語ベクトル $\boldsymbol{x}$、ターゲット単語ベクトル $\boldsymbol{y}$ の類似度を以下の式に従って計算する。

$$CSLS(\boldsymbol{x}, \boldsymbol{y}) = \cos(\boldsymbol{x}, \boldsymbol{y}) - mnn_T(\boldsymbol{x}) - mnn_S(\boldsymbol{y}) \tag{2.10}$$

ここでコサイン類似度 $\cos(\boldsymbol{x}, \boldsymbol{y})$ は以下の式によって定義される。

$$\cos(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{|\boldsymbol{x}||\boldsymbol{y}|} \tag{2.11}$$

また、$mnn_T(x)$ は $\boldsymbol{x}$ に類似度が高い上位 K 個のターゲット単語ベクトルとのコサイン類似度の平均であり以下の式によって表される。

$$mnn_T(\boldsymbol{x}) = \frac{1}{K} \sum_{\boldsymbol{y}_i \in N} \cos(\boldsymbol{x}, \boldsymbol{y}_i) \tag{2.12}$$

同様に $mnn_S(\boldsymbol{x})$ は $\boldsymbol{y}$ に類似度が高い上位 K 個のソース単語ベクトルとのコサイン類似度の平均である。

教師なし単語写像

近年では教師なしで線形写像を学習する手法も登場している。教師なしで写像を学習する手法では敵対的学習 [3] から作成した多言語埋め込み表現や単語の類似度分布を利用 [4] すること等によって初期の単語辞書を作成

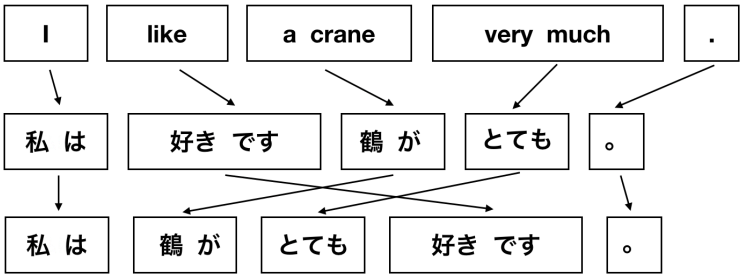

図 2.6. フレーズベース機械翻訳の例

し、その単語辞書を利用して再度式 (2.9) に従って変換行列を学習していく。同様にして単語辞書の生成と変換行列の学習を繰り返すことで変換行列を更新する手法が主に用いられている。

近年の研究動向

Søgaard ら [5] は異なる言語間にて別々のベクトル空間で訓練した単語埋め込み表現の幾何学的構造が必ずしも似ているわけではないこと示し、幾何学的構造が似るためには言語的に似た構造を持つ言語同士で共通するアルゴリズム、ドメインによる学習が必要であることを示した。

このような理由から線形写像のみで多言語埋め込み表現を学習するには限界があり、近年では対訳コーパスを利用する手法 [11] や機械翻訳の出力結果を利用する手法 [12] に対しても研究が進められている。本研究では教師なし機械翻訳の出力結果にて拡張した訓練データを用いて多言語埋め込み表現の学習を行う。

2.3　教師なしフレーズベース機械翻訳

フレーズベース機械翻訳

統計的機械翻訳においてターゲット文を y ソース文を x とすると、ターゲット文は次の式によって決められる。

$$\arg\max_{y} P(y|x) = P(x|y) \times P(y) \tag{2.13}$$

ここで $P(y)$ はターゲット文 y が出現する確率を表す。この確率はターゲット文がどれだけ尤もらしいかを表し、これをモデル化したものを言語モデルという。また $P(x|y)$ はソース文からあるターゲット文が出てくる条件付き確率を表し、これをモデル化したものを翻訳モデルと呼ぶ。統計的機械翻訳では言語モデルと翻訳モデルの性能を考慮して出力文を定める。

フレーズベース機械翻訳 (Phrase-Based Statistical Machine Translation, PBSMT) [13] は統計的機械翻訳の手法の 1 つであり連続した 1 単語以上の単語列であるフレーズごとの対応表であるフレーズテーブルからフレーズごとに翻訳し、自然な言語になるように並び替えを行うことで翻訳する。フレーズベース機械翻訳の学習ではプロの翻訳家が作成した対訳コーパスから学習するものや単語対応関係を利用する手法がある。フレーズベース機械翻訳の例を図 2.6 に示した。

教師なしフレーズベース機械翻訳

教師なし機械翻訳とは対訳コーパスを使わず、コンパラブルコーパスのみで機械翻訳を実現するシステムである。教師なし機械翻訳の手法の一つで、当時性能の大幅な向上に成功した Lample らの手法 [14] を中心に説明する。アルゴリズムの概要を **Algorithm 1** に記載した。

以下では翻訳される前の文をソース分、翻訳された文をターゲット文と呼び、ソース文からターゲット文への翻訳モデルを $P_{s \to t}$ などと書く。

Algorithm 1 Lample らの教師なしフレーズテーブル機械翻訳システムの学習アルゴリズム概要

```
それぞれの単言語コーパスから多言語埋め込み表現の作成

それぞれの単言語コーパスからソース言語モデル P_s の形成

P_s とフレーズテーブルを利用して最初の翻訳モデル P^0_{s→t} を作成

P^0_{s→t} と訓練データから擬似ターゲット文 D^0_t を生成

for k=1 to N do
    D^{k-1}_t と訓練データを擬似対訳コーパスとして翻訳モデル P^k_{t→s} を生成

    P^k_{t→s} と訓練データから D^k_s を生成

    D^k_s と訓練データを擬似対訳コーパスとして翻訳モデル P^k_{s→t} を生成

    P^k_{s→t} と訓練データから D^k_t を生成
```

コンパラブルコーパスに対して、単語埋め込み表現をそれぞれ学習させ、Conneau らの手法 [3] により多言語埋め込み表現を学習する*2。多言語埋め込み表現獲得後、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。

$$p(t_j|s_i) = \frac{\exp\left(\frac{1}{T}\cos(e(t_j), e(s_i))\right)}{\sum_k \exp\left(\frac{1}{T}\cos(e(t_k), e(s_i))\right)} \tag{2.14}$$

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、W はソース単語埋め込み表現空間をターゲット単語埋め込み表現空間に写像する回転行列を表し、$e(x)$ は x の多言語埋め込み表現*3を表す。

続いて n-gram 言語モデルである KenLM [16] を利用して言語モデルを作成する。

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 [17] を行うことで翻訳モデルの精度を向上させる。

2.4 BLEU スコア

機械翻訳において、翻訳の精度を自動で評価する手法として BLEU スコア (Bilingual Evaluation Understudy) [18] がある。これは機械翻訳の結果 (candidate) が翻訳者の翻訳の結果 (reference) に似ていれば似ているほど candidate の精度は高いだろうという前提に基づき作られた評価手法であり candidate と reference の類似度を以下のような式で計算する。

$$\mathrm{BLEU} = \mathrm{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{2.15}$$

ここで BP (brevity penalty) は candidate が短い時に BLEU が高くなってしまうのを防ぐために使われるペナルティで以下の式で計算される。

$$\mathrm{BP} = \begin{cases} \exp(1-\frac{r}{c}) & if\ c > r \\ 1 & (otherwise) \end{cases} \tag{2.16}$$

ここで c,r はそれぞれ、参照文のテストコーパスにおける系列長の総和を示す。また、p_n は n-gram の一致率を表し以下の式で計算される。

*2 Lample らは教師データとして対訳コーパスは一切利用していないが、多言語埋め込み表現の学習の際には単語辞書を用いている。

*3 2 単語以上を用いる場合はフレーズの埋め込み表現 [15] を利用する。

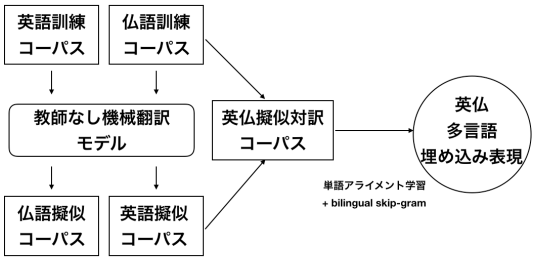

図 2.7. Marie らの手法概要

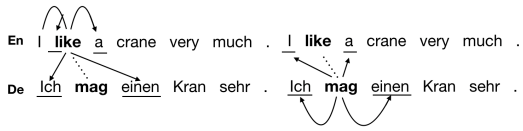

図 2.8. biligual skip-gram

$$BP = \begin{cases} \exp(1 - \frac{r}{c}) & \text{if } c > r \\ 1 & \text{(otherwise)} \end{cases} \tag{2.17}$$

w_n は一致率に対する適当な重みであり $\frac{1}{N}$ がよく使われる。

2.5 教師なし機械翻訳結果を利用した研究

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。

教師なし機械翻訳の結果から単語辞書の作成

Artetxe ら [19] は元の訓練コーパス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 FastAlign [20] にて単語の対応関係を学習させた後に、そこから 1 単語からなるフレーズテーブルを作成し単語辞書を取得している。作成した単語辞書について BLI にて既存手法より精度が高いことを示し、写像手法以外にも単語辞書を取得する方法の可能性を示唆した。

教師なし機械翻訳の結果に bilingual skip-gram

Marie ら [12] は元の訓練コーパス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 skip-gram [11] を適用することで多言語埋め込み表現を学習した。モデルの概要を図 2.7 に示す。bilingual skip-gram は単語対応の取れた対訳コーパスに対して、skip-gram アルゴリズムにより学習する。すなわち図 2.8 のように、ソース文のある注目単語についてその周囲の単語だけではなく、対応するターゲット文の単語の周辺単語も同時に推測することで、注目単語の多言語埋め込み表現を学習する。bilingual skip-gram による学習は対訳コーパスの単語対応の精度が多少弱くてもノイズに頑健な多言語埋め込み表現を作成できることが報告されている。Marie らは、提案した手法による多言語埋め込み表現の BLI が既存手法よりも高いことを示した。また、BLEU スコアのより高い教師なし機械翻訳による擬似対訳コーパスによる多言語埋め込み表現の方が BLI の精度がより高くなること、擬似対訳コーパスはソース言語とターゲット言語による擬似コーパスを両方学習に用いると逆に BLI が下がることを示した。さらに、Marie らの手法による多言語埋め込み表現を用いて教師なしフレーズベース機械翻訳を作成する場合、初期の翻訳モデルの BLEU スコアは既存手法より優れているが、逆翻訳を繰り返した際の翻訳モデルの性能は既存手法よりよくならないことを報告している。

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コーパスの拡張コーパスとして連結し、単語埋め込み表現の学習と写像により再度多言語埋め込み表現を学習する手法を提案する。

第 3 章

教師なし機械翻訳機から作成した多言語埋め込み表現の分析

3.1 提案手法

図 3.1 に本提案モデルによる多言語埋め込み表現獲得手法概要を示す。

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を訓練コーパスとして、単語埋め込み表現を獲得し、Artetxe らの教師なしの単語写像 [4] により、各単語埋め込み表現を同一空間上に写像することで多言語埋め込み表現を獲得する。多言語埋め込み表現を利用することで Lample らの手法 [14] により教師なしフレーズテーブル機械翻訳機[*1]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。

多言語埋め込み表現の再獲得

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、再度同じ手順で多言語埋め込み表現を学習する。Lample ら [14] の設定では教師データとして MUSE[*2]の単語辞書を用いて多言語埋め込み表現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。

3.2 実験設定

コーパス

ソース言語を英語、ターゲット言語をフランス語、ドイツ語、日本語として実験を行った。各言語、訓練コーパスは Wikipedia Comparable Corpora[*3]を利用し、それぞれ 1000 万文を用いた。また、英仏、英独テスト用の対訳コーパステスト用対訳コーパスとして WMT14[*4]の newstest2014 を用い、英日テスト用の対訳コーパスとして田中コーパス[*5]を用いた。

英語、フランス語、ドイツ語に対しては全て小文字化した後、mosesdecoder[*6]により単語分割を行った。日本

[*1] (1) 式において T=30 に設定し、ソース言語における頻出語 30,000 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。
[*2] https://github.com/facebookresearch/MUSE
[*3] https://linguatools.org/tools/corpora/wikipedia-comparable-corpora/
[*4] http://www.statmt.org/wmt14/translation-task.html
[*5] http://www.edrdg.org/wiki/index.php/Tanaka_Corpus
[*6] https://github.com/moses-smt/mosesdecoder

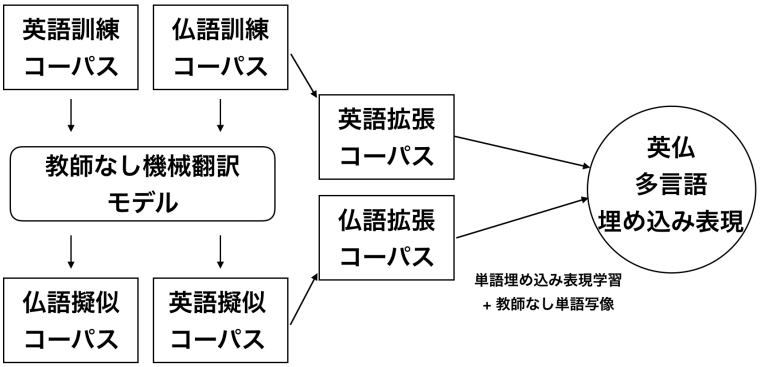

図 3.1. 提案手法概要

逆翻訳回数	en→fr	fr→en	en→de	de→en	en→ja	ja→en
0		14.5		10.5	1.2	
1	16.7	18.8	9.3	13.3	2.9	1.5
2	19.0	19.1	10.3	13.7	**3.7**	**1.5**
3	**19.3**	**19.0**	**10.3**	**13.7**	3.6	1.4

表 3.1. 教師なし機械翻訳モデルの BLEU スコア

表中の太字の BLEU スコアを持つ翻訳モデルを提案手法で利用した。

語に対しては kytea[*7]を用いて、単語分割を行った後、ユニコード正規化を行い、mojimoji[*8] にてアルファベットと数値を全て半角に、カタカナは全て全角に変換した。

教師なし機械翻訳モデル

fastText[*9]により Lample らの実験に倣い 512 次元、ウインドウサイズ 5、負例サンプリング 5 で単語埋め込み表現を学習した。それぞれの言語ペアに対して VecMap[*10]により多言語埋め込み表現を学習した。

Lample らが公開している実装[*11] を用いて教師なしフレーズベース機械翻訳モデルを作成した。(2.14) 式において T=45 に設定し、ソース言語における頻出語 30,000 語それぞれに対して (2.14) 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。

その後アルゴリズム 1 の手順に従って訓練データからイテレーションごとにランダムで 10 万文取得し、翻訳モデルからの擬似対訳コーパスの生成と mosesdecoder を用いた教師ありフレーズベース機械翻訳の学習 [13] を行うことで翻訳モデルの精度を向上させた。

擬似コーパスがどの程度の質である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逆翻訳の際の BLEU スコアを算出した。BLUE スコアの算出には mosesdecoder の実装を利用している。BLUE スコアの結果を表 3.1 に示す。表 3.1 から各言語の翻訳モデルにおいて逆翻訳により BLEU スコアが向上していることがわかる。また、言語の性質が比較的遠い英語・日本語間の BLEU スコアは低いことがわかる。

[*7] http://www.phontron.com/kytea/index-ja.html
[*8] https://github.com/studio-ousia/mojimoji
[*9] https://github.com/facebookresearch/fastText
[*10] https://github.com/artetxem/vecmap
[*11] https://github.com/facebookresearch/UnsupervisedMT

	en→fr	fr←en	en → de	de← en	en→ ja	ja←en
訓練コーパスのみ	0.664	0.636	0.561	0.567	0.451	0.357
bivec	0.620	0.594	0.527	0.520	0.263	0.274
提案手法	**0.696**	**0.669**	**0.637**	**0.612**	**0.488**	**0.418**

表 3.2. 既存手法との BLI の比較

l1→l2 は言語 l1 をソース言語、l2 をターゲット言語として BLI を計算することを表す。提案手法はターゲット言語のみを擬似コーパスで拡張した手法を示している。

全ての実験において写像を平等にするために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。

3.3 Bilingual Lexicon Induction による評価

3.3.1 既存手法との比較

本節では既存手法と提案手法について BLI での性能を比較した。テストデータは英語・ドイツ語、英語・フランス語は xling-eval [21] を用い、英語・日本語は Google Translate[*12]から構築したものを用いた。既存手法として以下の手法と比較した。

- 訓練コーパスのみを利用して単語埋め込み表現学習の後、教師なし写像により作成した多言語埋め込み表現 (訓練コーパスのみ)
- 2.5 節で述べた教師なし機械翻訳の結果に bilingual skip-gram を適用させることより作成した多言語埋め込み表現 (bivec)

前者の設定では訓練コーパスのみを用いて提案手法と同じ設定で fastText により 512 次元の単語埋め込み表現の学習を行い、VecMap による写像を行うことで多言語埋め込み表現を取得する。後者の設定では訓練コーパス、教師なしフレーズベース翻訳モデルは提案手法と同様のものを用いた。それ以外の設定は Marie らの実験に倣い、全てのソース訓練コーパスとそれにより出力した擬似コーパスを擬似対訳コーパスとして、試行回数 5 回、α=1、β=1 の設定で FastAlign [20] により単語対応関係を学習し、ウインドウサイズ 10、負例サンプリング 30 の設定で bivec[*13]により多言語埋め込み表現を学習した。

結果を表 3.2 に示す。結果からいずれの既存手法よりも BLI が向上していることがわかる。また、Marie ら [12] は本実験で用いた訓練コーパス量よりもさらに膨大な量のコーパスを用い、BLEU スコアのより高い教師なし機械翻訳モデルを利用して学習した多言語埋め込み表現にて BLI が向上したことを報告しているが、本実験の設定では BLI の性能が下がってしまうことが確認された。

また、実際に英語・日本語間の多言語埋め込み表現にて日本語の ‘レモン’ と英語の ‘papa’ という単語に対してそれぞれコサイン類似度が上位の写像先の言語の単語を 5 個表示することで順位が改善された例を表 3.3、3.4 に示す。

	1st	2nd	3rd	4th	5th
訓練コーパスのみ	juice	caramel	**lemon**	lemons	flavored
提案手法	**lemon**	juice	grapefluit	lemons	flavored

表 3.3. ‘レモン’ の近傍単語の改善例

[*12] https://translate.google.com/
[*13] https://github.com/lmthang/bivec

	1st	2nd	3rd	4th	5th
訓練コーパスのみ	フー	リトル	**パパ**	エルヴィス	プリティ
提案手法	**パパ**	ゴー	ママ	サンタマリア	ジョニー

表 3.4. ‘papa’ の近傍単語の改善例

赤文字は単語辞書テストデータにある翻訳先単語を表す

コーパス	fr	fr + psfr	de	de + psde	ja	ja + psja
en	0.655 ± 0.003	0.685 ± 0.001	0.548 ± 0.004	$\mathbf{0.615} \pm 0.002$	0.451 ± 0.002	$\mathbf{0.488} \pm 0.001$
en + psen(fr)	$\mathbf{0.704} \pm 0.001$	$\mathbf{0.693} \pm 0.000$	0.561 ± 0.005	0.614 ± 0.004	0.448 ± 0.002	0.487 ± 0.003
en + psen(de)	0.647 ± 0.004	0.680 ± 0.002	$\mathbf{0.610} \pm 0.004$	0.611 ± 0.004	0.445 ± 0.005	0.486 ± 0.006
en + psen(ja)	0.598 ± 0.002	0.652 ± 0.003	0.496 ± 0.005	0.588 ± 0.001	$\mathbf{0.463} \pm 0.002$	0.473 ± 0.002

表 3.5. 各コーパスペアに対する BLI の結果

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。

3.3.2 翻訳元の性質を反映している可能性の検証

前節では提案手法における BLI が既存手法に比べ向上することが確認されたが、その理由として翻訳文が翻訳元の言語の性質を反映していることが考えられる。Toral ら [22] は翻訳された文章は翻訳先の言語の実際の文章と比較すると、より標準化（特定の単語に限定）され、翻訳元の言語の性質を反映することがあると報告している。そこで本研究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単語埋め込み空間の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た。これを検証するため、写像先ではない言語から拡張する実験も行った。

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、以下のソース・ターゲットペアに対して 3 回異なる乱数シードで多言語埋め込み表現を学習し BLI での性能の平均と標準偏差を算出した。結果を表 3.5 に示す。

- 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 or 日本語）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 or 日本語） 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 or 日本語） 訓練コーパス ＋ 英語から生成した擬似コーパス

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が、ソース訓練コーパスを多言語埋め込み表現の写像先ではない、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。

また、ソース・ターゲット言語を両側拡張した場合よりも片側を拡張した場合の方が精度が高くなる傾向にあることが読み取れる。これはターゲット言語からの翻訳文が標準化され、ターゲット言語の性質を反映したとしても、ソース言語からの翻訳文の標準化のされ方、性質の保持の仕方とは同様であるとは限らず、かえって各言語の単語埋め込み表現学習の際の単語の共起情報が異なってしまうことが原因だと考えられる。

3.4 言語横断タスク

多言語埋め込み表現を評価する手法としては BLI が主に用いられてきたが Glavas ら [21] は BLI で高い精度を持つ多言語埋め込み表現が必ずしも言語横断的なタスクで良い性能を発揮するわけではないことを示している。そこで本実験では言語横断タスクである文書分類タスク、感情分析タスクにて多言語埋め込み表現の性能を検証した。

言語横断タスクではソース言語の訓練データを多言語埋め込み表現に変換し、その多言語埋め込み表現を入力として言語横断タスクを解く分類器を訓練する。ただしこの際、分類器のパラメータは更新させるが多言語埋め込み表現の値は更新させない。次に、分類器のパラメータは固定したままターゲット言語のテストデータを多言語埋め込み表現に変換し、その多言語埋め込み表現を入力として分類器の精度を測る。この精度が高いほど、多言語埋め込み表現上でより良く意味を共有できていると言える。

文書分類器として一層畳み込み層と max プーリング層とソフトマックス層で構成された畳み込みニューラルネットワーク (CNN) を利用した。畳み込み層の設定は Glavas ら [21] の実験に倣いフィルターサイズを 2〜5、フィルター数を 8 とした。

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、日本語をテスト、開発コーパスとして 20 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

文書分類タスクは与えられた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[*14]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 3.6 に示す。

感情分析タスク

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[*15]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 3.7 に示す。

カテゴリ	訓練	テスト (fr/de/ja)	開発 (fr/de/ja)
CCAT	8761	3996/3936/4084	1218/1276/1233
ECAT	8785	3892/4104/3816	1251/1253/1098
GCAT	8799	3992/4088/3924	1280/1214/1099
MCAT	8656	4120/3872/4176	1249/1256/1085

表 3.6. 文書分類に用いたデータセットの文数

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

表 3.7. 感情分析に用いたデータセットの文数

各言語、上記の文数を用いている

	en-fr	en-de	en-ja
訓練コーパスのみ	79.5 ± 1.5 (92.6)	79.0 ± 1.5 (91.7)	70.4 ± 1.2 (92.2)
提案手法	**82.2**† ± 1.5 (93.3)	79.3 ± 2.0 (92.0)	**71.6**† ± 0.8 (93.3)

表 3.8. 文書分類タスクの結果

[*14] https://github.com/facebookresearch/MLDoc
[*15] https://webis.de/data/webis-cls-10.html

	en-fr	en-de	en-ja
訓練コーパスのみ	69.1 ± 1.1 (71.8)	63.7 ± 1.3 (71.1)	63.5 ± 1.1 (70.7)
提案手法	69.5 ± 0.8 (71.9)	**65.1**† ± 0.8 (70.2)	62.8 ± 1.4 (70.6)

表 3.9. 感情分析タスクの結果

提案手法はターゲット言語のみを擬似コーパスで拡張した手法を示している。
† は既存手法と有意差があること（p < 0.01）を表し、括弧内の数字はソース言語（英語）でのテスト精度の平均を表す。

文章分類タスク、感情分析タスクの結果を表 3.8、3.9 に示す。文章分類タスクでは英語・フランス語間、英語・日本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。感情分析タスクでは英語・ドイツ語間にて既存手法のスコアを上回ることが確認されたが、英語・フランス語間、英語・日本語間では有意差は認められなかった。これらの結果から、提案手法による多言語埋め込み表現は文書分類タスクの性能において既存手法を上回る可能性が確認されたが、感情分析タスクでは一貫した傾向が得られなかった。本手法により単語の共起情報が言語間で似通ったとしても極性を持つ、もしくは対義語関係にある単語同士は似たような文脈で出現することが多いためそれらの差別化にはあまり寄与しないことが原因だと考えられる。

3.5 意味的類似度タスク

提案手法では教師なし機械翻訳による出力結果を学習データとして用いたが、翻訳文は必ずしも正確ではなく、様々なノイズが混入している。そこで本実験では擬似コーパスにより訓練コーパスを拡張して学習した単語埋め込み表現の質を意味的類似度タスク (Word Similarity) により検証した。

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係数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[*16] 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[*17] を用いた。

コーパス	simverb-3500	men
訓練コーパスのみ	0.259 ± 0.006	0.763 ± 0.001
擬似コーパスあり (fr)	0.260 ± 0.004	**0.767**† ± 0.002
擬似コーパスあり (de)	0.253 ± 0.003	**0.768**† ± 0.002
擬似コーパスあり (ja)	0.220† ± 0.001	0.760† ± 0.002

表 3.10. 意味的類似度タスクの結果

† は既存手法と有意差があること（p < 0.01）を表す。

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

表 3.11. 意味的類似度タスク（men）の結果の例

sim は人手による類似度、cos_sim はコサイン類似度を表す。

意味的類似度タスクの結果を表 3.10 に、結果の例を表 3.11 に示す。英語と言語的に遠い日本語においては simverb-3500、men のいずれのタスクにおいても精度が減少してしまっているが、比較的英語と言語的に近いフランス語、ドイツ語からの拡張では simverb-3500 では有意差は認められなかったが、men では提案手法のスコアが既存手法を若干上回ることが確認された。この結果から関係性の近い言語の擬似コーパスによる拡張はノイズが入ってるにも関わらず、単言語埋め込み表現自体の質を下げることがないことが示された。

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

逆翻訳回数	en → fr	fr → en	en → fr 提案手法	fr → en 提案手法
0		14.7		**14.8**
1	**16.7**	**18.8**	16.1	18.2
2	**18.8**	**19.2**	18.19	18.51
3	**19.2**	**19.1**	18.6	18.8

表 3.12. 教師なし機械翻訳機の BLEU スコア（en-fr）

逆翻訳回数	en → de	de → en	en → de 提案手法	de → en 提案手法
0		10.7		**10.9**
1	**9.6**	**13.5**	9.2	13.0
2	**10.3**	**13.5**	9.8	13.2
3	**10.2**	**13.5**	9.9	13.3

表 3.13. 教師なし機械翻訳モデルの BLEU スコア（en-de）

逆翻訳回数	en → ja	ja → en	en → ja 提案手法	ja → en 提案手法
0	**1.52**		1.08	
1	**3.55**	**1.56**	2.89	1.31
2	**3.58**	**1.55**	3.07	1.29
3	**3.42**	**1.47**	3.27	1.26

表 3.14. 教師なし機械翻訳モデルの BLEU スコア（en-ja）

なお、提案手法はターゲット言語のみを擬似コーパスで拡張した手法を示している。

3.6　教師なし機械翻訳への再利用

今回作成した多言語埋め込み表現で再度教師なし機械翻訳の訓練をすることで翻訳の精度が改善できるのかどうかを検証した。すなわち、提案手法による多言語埋め込み表現にて 3.2 節の教師なし機械翻訳モデルの作成と同様に翻訳モデルを作成し、各逆翻訳試行時の BLEU スコアを計測した。結果を表 3.12、3.13、3.14 に示す。

結果からフレーズテーブルと言語モデルにて作成した最初の翻訳モデルの BLEU スコアは提案手法と既存手法であまり変わらないが、その後逆翻訳を繰り返すと提案手法での BLUE スコアの向上率が悪く既存手法に逆転されてしまうことが確認された。

これらの結果は Marie らの作成した多言語埋め込み表現においても同様の傾向にあることが報告されている [12]。Edunov らの逆翻訳の実験 [23] では擬似コーパスの多様性が高いほど（パープレキシティが高いほど）翻訳モデルの学習を困難にし，ノイズに対して頑健な学習ができることを示している。3.3 節で述べたように翻訳文での語彙はある程度標準化される（語彙数が制限される）。実際に翻訳文の語彙数が制限されているのか確かめるために本実験で使用した訓練コーパスと擬似コーパスの単語数あたりの語彙数を比較した。その結果を表 3.15 に示す。結果からどの言語の擬似コーパスも訓練コーパスに比べて単語数あたりの語彙数が少ないことが確認された。そのため、擬似コーパスを利用する多言語埋め込み表現では特定の単語が写像されやすくなり、作成された翻訳モデルではより決まったパターンでフレーズが訳されやすくなる。その結果、翻訳される文章の多様性が低くなることで逆翻訳時の翻訳モデルの学習が容易になり、逆翻訳による BLEU スコアがあまり向上しなかったのだと

	en-fr		en-de		en-ja	
	en	fr	en	de	en	ja
訓練コーパス	1.60×10^{-3}	1.63×10^{-3}	1.51×10^{-3}	3.78×10^{-3}	1.52×10^{-3}	1.03×10^{-3}
擬似コーパス	0.57×10^{-3}	0.57×10^{-3}	0.66×10^{-3}	0.59×10^{-3}	0.19×10^{-3}	0.17×10^{-3}

表 3.15. 単語数あたりの語彙数の比較

考えられる。この結果は BLI の改善が必ずしも教師なし機械翻訳の精度改善には繋がらないことを示唆している。

第 4 章

おわりに

4.1　本研究のまとめ

本研究では擬似コーパスによってコンパラブルコーパスを拡張することで教師なしで作成された多言語埋め込み表現の性能をさらに向上させる手法を提案した。翻訳元の性質を反映した文章による単語埋め込み表現の学習は各単語埋め込み空間の幾何学構造を似通わせ写像精度を向上することができることを確認した。さらに、提案手法による多言語埋め込み表現は一部の言語横断タスクでも精度を向上させる可能性を示し、関係性が強い言語間に関しては擬似コーパスを利用しているにもかかわらず、単言語埋め込み表現自体の質を下げないことを示した。また、提案手法による多言語埋め込み表現により作成した教師なし機械翻訳モデルは出力文の多様性が低く、逆翻訳での BLEU 向上率が悪いことを示した。この結果は BLI の改善が必ずしも教師なし機械翻訳の精度改善には繋がらないことを示唆している。

4.2　今後の展望

今後の展望としては提案手法による多言語埋め込み表現に対してさらなる試行回数や異なる言語による実験を行うことで有用性を検証する必要がある。

また、本実験では言語横断タスクとして文書分類タスク、感情分析タスクにより評価を行ったが、言語横断タスクの種類によって BLI との相関が異なることが報告されている [21]。よって今後は言語横断的な情報抽出や係り受け解析など、他のタスクでの性能評価 [1] が考えられる。

さらに本研究では教師なしでコンパラブルコーパスを拡張する手法を提示したが、対訳コーパスを用いて教師ありで訓練されたより翻訳性能の高い機械翻訳モデルの出力結果から作成した多言語埋め込み表現の分析を考えている。

参考文献

[1] Sebastian Ruder, Ivan Vuliundefined, and Anders Søgaard. A survey of cross-lingual word embedding models. *J. Artif. Int. Res.*, Vol. 65, No. 1, p. 569–630, May 2019.

[2] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations*, 2013.

[3] Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. In *International Conference on Learning Representations*, 2018.

[4] 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*, pp. 789–798, May 2018.

[5] Anders Søgaard, Sebastian Ruder, and Ivan Vuli. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 778–788, Melbourne, Australia, July 2018.

[6] Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*, 2018.

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

[8] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, Vol. 5, pp. 135–146, 2017.

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

[10] Miloš Radovanoviundefined, Alexandros Nanopoulos, and Mirjana Ivanoviundefined. Hubs in space: Popular nearest neighbors in high-dimensional data. *J. Mach. Learn. Res.*, Vol. 11, p. 2487–2531, December 2010.

[11] Thang Luong, Hieu Pham, and Christopher D. Manning. Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pp. 151–159, Denver, Colorado, June 2015.

[12] Benjamin Marie and Atsushi Fujita. Unsupervised joint training of bilingual word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3224–3230, Florence, Italy, July 2019.

[13] Philipp Koehn, Franz J. Och, and Daniel Marcu. Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 127–133, 2003.

[14] Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc Aurelio Ranzato. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 5039–5049, Brussels, Belgium, October-November 2018.

[15] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734, Doha, Qatar, October 2014.

[16] Kenneth Heafield. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pp. 187–197, Edinburgh, Scotland, July 2011.

[17] Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, August 2016.

[18] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA.*, pp. 311–318, 2002.

[19] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Bilingual lexicon induction through unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5002–5007, Florence, Italy, July 2019.

[20] Chris Dyer, Victor Chahuneau, and Noah A. Smith. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 644–648, Atlanta, Georgia, June 2013.

[21] Goran Glavaš, Robert Litschko, Sebastian Ruder, and Ivan Vuli. How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 710–721, Florence, Italy, July 2019.

[22] Antonio Toral. Post-editese: an exacerbated translationese. *the 17th Machine Translation Summit.*, 2019.

[23] Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 489–500, Brussels, Belgium, October-November 2018.

[24] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to Sequence Learning with Neural Networks. In *Advances in Neural Information Processing Systems*, pp. 3104–3112, 2014.

[25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.

[26] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, pp. 1096–1103, 2008.

[27] G E Hinton and R R Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, Vol. 313, No. 5786, pp. 504–507, jul 2006.

[28] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, Vol. 313, No. 5786, pp. 504–507, 2006.

謝辞

本研究を進めるにあたって、多くの皆様にお世話になりました。

指導教員の鶴岡慶雅教授には毎回のミーティングにて研究に対するアドバイスをいただきました。また、論文、発表の資料についても丁寧に添削をしていただきました。

修士 2 年の李凌寒さんには、論文の添削に加えて研究の方針や実験設定について相談にのっていただきました。

研究室の先輩方には、研究についてのアドバイスだけでなく、論文の読み方・探し方、研究に用いるサーバーの利用方法など研究の進め方について幅広く教えていただきました。

この場をお借りして、お世話になった皆様に厚く御礼申し上げます。