Vol.2019-NL-243 No.9
2019/12/4

Encoder-Decoderモデルを用いた対訳コーパスからのマルチリンガル単語分散表現の獲得

和田 崇史[1,a)]　岩田 具治[2,3,b)]　松本 裕治[1,3,c)]

概要：本研究は, 対訳コーパスを用いてマルチリンガル単語分散表現を学習する新たな手法を提案する. 多言語で翻訳および入力文の生成を行う Encoder-Decoder モデルを学習することで, 全言語共通の空間で表されたマルチリンガル単語分散表現を獲得する. 実験では対訳辞書生成のタスク (Billingual Lexicon Induction) で単語分散表現の精度を評価し, 提案手法が既存手法を上回っていることを示した.

1. はじめに

異なる言語の単語と単語の対応関係を明らかにする手段として, マルチリンガル単語分散表現 (Multilingual Word Embedding) の研究が近年盛んに行われている. マルチリンガル単語分散表現は複数言語の単語を同じベクトル空間で表す研究で, 機械翻訳やクロスリンガル情報抽出等の複数言語を扱うタスクに応用されている [1], [2], [3]. また, 高資源言語から低資源言語への転移学習 (Transfer Learning) の手法としても有効的である [4].

マルチリンガル単語分散表現はこれまでに様々な手法が提案されているが, 大きく分けると 2 種類のアプローチに分類される. 一つは word2vec [5] などで各言語のコーパスから独立に学習された単語分散表現を共通の空間に写像する「マッピング」の手法, もう一つは対訳コーパス等を用いて単語分散表現を同じ空間で同時に学習する「同時学習」の手法である. マッピングは資源豊富な単言語コーパスを活用できるというメリットがあるが, 「異なる言語の分散表現が線形に写像可能」という条件を仮定するため, アラインする言語の文法が大きく異なる場合や資源が少ない状況では学習が困難だと知られている [6], [7]. 一方, 同時学習はそのような条件を必要としないため, 対訳コーパスが存在する場合においては有望なアプローチである. 本研究は後者に属する手法を提案する.

対訳コーパスを用いる同時学習の主な既存手法は, 文のアラインメント情報を考慮しながら Skip Gram [5] を用いて単語分散表現を学習する. Skip Gram はターゲット単語の分散表現から文脈 (Context) を予測する手法で, 既存手法は言語非依存な文脈を対訳コーパスから抽出しマルチリンガルな表現を学習する. 例えば, 先行研究 [8] では対訳コーパスでアラインされた文のセットに共通の文 ID を割り当て, その ID を単語の「文脈」として予測する. また, 対訳コーパスから単語とフレーズのアラインメントを抽出し, それらをお互いに予測する手法 [9], [10] や, アラインされた単語の周辺単語を文脈として学習する手法 [11] がある.

しかし, これらの手法では語順を（一部）無視して学習するため, 文の系列情報を十分に考慮できていないという問題がある. そこで, 本研究は文の系列情報を十分に活用するため, 翻訳と入力文の生成を多言語で行う Encoder-Decoder モデルを学習し, マルチリンガルな単語分散表現を同時に学習する. 評価には対訳辞書生成（Billingual Lexicon Induction）の実験を行い, 提案手法が既存手法の精度を大幅に上回ることを示した. また, 対訳データを用いない教師なしの設定においても, 提案手法が効果的であることを示した.

2. 提案手法

2.1 マルチリンガル単語分散表現

本研究はマルチリンガル単語分散表現の獲得を行う. L 言語のコーパスが与えられた時, 言語 $\ell(1 \leqq \ell \leqq L)$ の単語 $w^\ell \in V^\ell$ を全言語で共通の分散表現空間に埋め込み, 同じ意味を持つ単語間の距離が言語 ℓ に依らず近くなる表現を学習する. 本研究では, 異なる言語間で文のアラインメン

[1] 奈良先端科学技術大学院大学
Nara Institute of Science and Technology
[2] NTT コミュニケーション科学基礎研究所
NTT Communication Science Laboratories
[3] 理化学研究所革新知能統合研究センター
RIKEN Center for Advanced Intelligence Project (AIP)
[a)] wada.takashi.wp7@is.naist.jp
[b)] tomoharu.iwata.gy@hco.ntt.co.jp
[c)] matsu@is.naist.jp

ⓒ 2019 Information Processing Society of Japan

トが取れた対訳文コーパスを用いて学習を行う.

2.2 モデル概要

本提案手法は LSTM を用いた標準的な Encoder-Decoder モデルを用いており, 多くのパラメータを複数の言語でシェアしている. 具体的なパラメータは以下の通り.

- **言語 ℓ 毎に異なるパラメータ**
  - 単語分散表現: $\boldsymbol{E^{\ell}} \in \mathbb{R}^{\boldsymbol{D}\times|\boldsymbol{V^{\ell}}|}$
  - Decoder の LSTM: $\boldsymbol{g^{\ell}}$
- **全言語で共通のパラメータ**
  - Endoder の両方向 LSTM: $\overrightarrow{\boldsymbol{f}}$, $\overleftarrow{\boldsymbol{f}}$
  - 隠れ層次元 H を単語分散表現の次元 D に写像する行列: $\boldsymbol{W} \in \mathbb{R}^{\boldsymbol{H}\times\boldsymbol{D}}$

N 単語から成るソース言語 s の文 $\langle w_1^s ..., w_N^s \rangle$ から M 単語のターゲット言語 t の文 $\langle w_1^t ..., w_M^t \rangle$ を生成する場合を考える. Encoder は両方向 LSTM によって語順を考慮した文表現を獲得する.

$$\overrightarrow{u}_i = \overrightarrow{f}(\overrightarrow{u}_{i-1}, E^s\Psi(w_{i-1}^s)), \tag{1}$$

$$\overleftarrow{u}_i = \overleftarrow{f}(\overleftarrow{u}_{i+1}, E^s\Psi(w_{i+1}^s)), \tag{2}$$

$$u_i = [\overrightarrow{u}_i, \overleftarrow{u}_i] \tag{3}$$

ここで, $\Psi(w)$ は単語 w の one-hot ベクトル, $[x, y]$ は x と y の連結を表す. Encoder の情報を用いて, Decoder は文頭から文末の順向き生成と, その逆方向の生成を行う. 両方向の生成は共通の LSTM g^t で学習され, { 初期入力 w_0^t, 文末記号 w_{M+1}^t} のペアを {<BOS>, <EOS>}（順向き生成）または {<EOS>, <BOS>}（逆向き生成）と入れ変えて学習する. なお, 初期入力を変えることで NMT の生成の方向をコントロール出来る事はこれまでにも報告されている [12]. 両方向の生成は以下のように各方向で独立に学習される.

$$p(w_1^t ..., w_{M+1}^t) = \prod_{i=1}^{M+1} p(w_i^t|\overrightarrow{h}_i, u). \tag{4}$$

$$p(w_0^t, ..., w_M^t) = \prod_{i=0}^{M} p(w_{M-i}^t|\overleftarrow{h}_{M-i}, u). \tag{5}$$

$$\overrightarrow{h}_i = g^t(\overrightarrow{h}_{i-1}, E^t\Psi(w_{i-1}^t))), \tag{6}$$

$$\overleftarrow{h}_i = g^t(\overleftarrow{h}_{i+1}, E^t\Psi(w_{i+1}^t))). \tag{7}$$

通常, Decoder の初期値 $\overrightarrow{h}_0$ または $\overleftarrow{h}_{M+1}$ には Encoder の最終隠れ層を結合したベクトル $[\overrightarrow{u}_N, \overleftarrow{u}_0]$ を用いることが多いが, 本提案手法では言語モデルと同様に, 常に同じ初期値から生成させた. その結果, 単語分散表現の質は落とさずに計算時間とメモリー使用量を削減することができた[*1]. 生成単語の確率分布は以下のように求める.

$$p(w_i^t|h_i, u) = \mathrm{softmax}(E^{t\top} h_i') \tag{8}$$

$$h_i' = W(c_i + h_i), \tag{9}$$

$$c_i = \sum_{j=1}^{N} \alpha_{i,j}^t u_j, \tag{10}$$

$$\alpha_{i,j}^t = \frac{\exp(h_i u_j)}{\sum_{k=1}^{N} \exp(h_i u_k)}. \tag{11}$$

式 (8) において, 出力写像の重みを単語分散表現と共有した (Weight Tying [13]). その結果, 言語 ℓ に対して Encoder の入力層と Decoder の入出力層の 3 箇所で同じ単語表現 E^{ℓ} が用いられる (Three-Way Weight Tying [14]). これにより, 本研究の目的としている単語分散表現の学習が効率的に進むと期待される. また, 簡略化のため式 (9) では h_i' を Encoder の文脈ベクトル c_i と Decoder の隠れ層 h_i の和を線形に写像した[*2].

加えて, 各言語の単語表現が独立の空間で学習されることを防ぐために, 提案手法では単語分散表現に対して Dropout [16] を適応した. 具体的には, 式 (1,2,6,7,8) で用いる重み行列 E^s と E^t に対してそれぞれ Dropout を 0.5 の確率で適用した. これにより, 各言語の単語分散表現空間が分離することを直接抑えることができる[*3].

2.3 学習方法

本モデルは, 翻訳と入力文の生成を同時に学習する. 1 言語の入力文から複数の言語の対訳文と入力文自身を生成する学習により, マルチリンガルな文および単語の単語表現が獲得できる. L 言語間でアラインされた対訳コーパスが与えられた時, 以下の誤差を最小化するようにモデルを学習する.

$$E = \sum_{s=1}^{L}\sum_{t=1}^{L}\sum_{j=1}^{C_{st}}\sum_{d=1}^{2}\sum_{i=1}^{M_j} \Delta(y_{j,i}^t, p(y_{j,i}^t|h_{j,i}^d, u_j)). \tag{12}$$

ここで, C_{st} は対訳文数, M_j はターゲット文 y_j の単語数, d は生成方向 ($h^1, h^2 = \overrightarrow{h}, \overleftarrow{h}$) を表す. なお, 損失関数 Δ には交差エントロピーを用いた. 上式が表すように, 本提案手法は学習言語数 L に対して $\mathcal{O}(L^2)$ の計算時間がかかりコストが高いという欠点がある. そこで, 言語 s のミニバッチサイズ B の文を Encoder で読み込んだ後, Decoder で LB 文を生成し, 一度に損失を計算してパラメータを更新した. これにより, 文を入力する回数とパラメータ更新回数を各生成方向で $\frac{1}{L}$ 倍に減らすことができ, 計算量削減となる.

[*1] 文生成の perplexity は少し上昇した

[*2] 通常は非線形関数を用いて $h_i' = W_{out}(\tanh([c_i, h_i]))$ 等で求めることが多い [15]

[*3] スタンダードな機械翻訳モデルでは, Dropout は隠れ層 h_i 及び h_i' に適用することが多い

ⓒ 2019 Information Processing Society of Japan

3. 実験

3.1 データと実験設定

本研究では聖書対訳コーパス (Parallel Bible Corpus; PBC)[17] でベースラインおよび提案手法の学習を行った. PBC は全言語で文のアラインメントが取れている多言語パラレルコーパスである. 本研究は学習に各言語で 22,456 文を用い , モデル選択を行う開発データとして 1,000 文を用いた. 前処理として, tokenize した後に全ての文字を小文字化した. tokenize は UDPipe [18] の ver. 2.4 で行なった. 実験は以下の 3 つの言語セットで行なった.

- **実験 1：スペイン語 (es), 英語 (en)**
- **実験 2：スペイン語 (es), フランス語 (fr), イタリア語 (it), 英語 (en)**
- **実験 3：日本語 (ja), 英語 (en)**

実験 1,2 は基本語順が主語-動詞-目的語（SVO）で共通しているため, 比較的アラインメントが取りやすい言語セットだと考えられる. 一方, 実験 3 は基本語順が SOV と SVO で異なるため, アラインメントの難易度が高い言語セットだと言える. これら 3 つのセットでマルチリンガル単語分散表現をそれぞれ学習し, その精度を以下の実験で評価した.

3.2 評価方法

本研究では対訳辞書生成 (Billingual Lexicon Induction; BLI) の実験を行なった. Q 対から成るソース言語 s とターゲット言語 t の対訳辞書 (w_i^s-$w_i^t, i = 1, 2..., Q$) が与えられた時, ソース言語の各単語 w_i^s に最も類似度が高い単語をターゲット言語の全語彙 V^t から抽出し, アラインされた単語が実際に対訳ペアであるかどうかの正解率 (p@1) で評価する. 類似度の計算は先行研究 [19], [20] に倣って, 単語分散表現の Cross-domain Similarity Local Scaling (CSLS)[19] を用いた. CSLS はコサイン類似度をベースにした手法で, ハブ問題 (hubness problem) を緩和することができるため候補数が多いアラインメントで有効的な手法である. なお, 評価データには先行研究 [19] が公開した最大で 10 万のペアからなる対訳辞書を用い, そこから学習データの語彙に含まれる 1000 単語を各言語[*4]で抽出して評価に用いた.

3.3 ベースライン

ベースラインには, 各言語で事前学習された単語分散表現を共通の空間に写像する「マッピング」の手法と, 対訳コーパスを用いて単語分散表現を同時に学習する「同時学習」の手法を学習し, 提案手法と比較した. 単語分散表現の次元数は全ての手法で 500 に統一し, 使用頻度が 2 回以上の単語を語彙に含めた. これは, 同じく PBC で実験を行なった先行研究 [8] と同じ実験設定である.

3.3.1 マッピング

マッピングのベースラインには教師なし手法の VecMap [20] を用いた[*5]. VecMap は 2 言語間で写像の学習と擬似対訳辞書の生成を繰り返す自己学習をベースとしたバイリンガルな手法で, 最も精度の高い教師なし手法の一つである [24]. さらに, 対訳辞書を用いる教師ありの手法とも比較した. データは, [19] の対英辞書から評価に用いた単語を取り除いて残った全ての単語ペアを用いた[*6]. 手法は, Procrustes Problem [21] を解く基本的な手法 (5.1 参照) と, それをさらに発展させた RCSLS [22] を用いた. いずれの場合も, マッピングする各言語の単語分散表現学習は PBC 対訳コーパスに対して FastText [25] を用いて事前学習した. なお, 全ての実験において英語以外の言語の単語分散表現を英語の空間に写像した.

3.3.2 同時学習

提案手法と同じアプローチである教師あり同時学習のベースラインとして, 文 ID で Skip Gram を行う手法 [8]（以下 S-ID）と XLM [23] と比較した. S-ID は, 対訳コーパスの各文に言語非依存の ID を割り当て各単語から文の ID を予測する手法で, 同時学習で最も精度が高い手法の一つである [8], [10]. 後者の XLM は複数言語で共通の BERT [26] を学習してマルチリンガルな表現を獲得する手法である. XLM は本来, Wikipedia のような大規模のデータを用いて学習するモデルで, また Byte Pair Encoding (BPE) [27] でサブワード化することを前提としているため, マルチリンガルな単語分散表現を得る目的では通常用いられない. しかし, 本研究と同じく語順を考慮した手法であるため, 比較のため BPE を適用せずに XLM を学習して, 提案手法と比較した. 学習方法は, 全ての言語の組み合わせで Masked Language Model (MLM) と Translation Language Model (TML) を聖書コーパスのみで学習した（e.g., 実験 2 の場合, 4 言語の MLM と 6 言語ペアの TML を同時に学習した）. なお, XLM のパラメータは提案手法と同程度の数に設定した [*7].

また, 教師なしの手法である MNLM [6] とも比較した. アラインメント情報を失わせるために PBC コーパスを各言語で独立にシャッフルし, それぞれ単言語コーパスとして扱った. 最大で 50 エポック学習し, 損失が飽和（損失の改善が前エポックと比べて 0.5%以下）した時に学習を止めた. また, MNLM に本提案手法を適用した場合の精度も

[*4] 日本語は 803 単語のみしか存在しなかったので, それらを全て用いた

[*5] 全言語を同じ空間に写像するため, 単語分散表現の re-weighting, whitening, 及び normalisation のプロセスは省いた

[*6] 単語語数: スペイン語 3218, フランス語 2639, イタリア語 2567. 日本語は十分なデータが存在しなかったため, 行わなかった.

[*7] 隠れ層次元 500, MLP 次元 2000 の 2 層の Transformer, multi head 数は 4

ⓒ 2019 Information Processing Society of Japan

学習方法	教師なし			教師あり				
	マッピング	同時学習		マッピング		同時学習		
手法 / 言語ペア	VecMap [20]	MNLM [6]	OURS (LM)	Procrustes [21]	RCSLS [22]	XLM [23]	S-ID [8]	OURS
ja-en (bi)	0.9	0.0	4.1	–	–	0.0	39.1	**47.2**
es-en (bi)	0.2	20.8	16.9	13.6	14.8	5.8	53.7	**65.3**
es-en	0.2	23.1	25.9	13.6	14.8	7.2	54.7	**65.3**
fr-en	0.3	20.0	21.7	13.2	14.6	12.1	47.4	**58.3**
it-en	0.0	17.1	21.3	7.5	8.5	6.7	49.4	**59.7**
es-fr	0.0	28.7	36.7	15.8	16.5	10.7	63.2	**70.9**
es-it	0.0	29.9	36.8	18.5	16.8	19.5	60.7	**68.7**
fr-it	0.0	31.7	38.7	17.1	16.7	8.7	63.3	**69.9**

表 1: 対訳辞書生成タスクにおける, 実験 1, 2, 3 での各手法の正解率 (p@1 %). (bi) は 2 言語 (bilingual) の対訳コーパスでモデルを学習したことを意味し, es-en (bi) と ja-en (bi) はそれぞれ実験 1, 3 を示す.

比較した (**OURS (LM)**). これは, 提案手法から注意機構と Encoder を省き, MNLM と同様に全言語で共通の言語モデルを学習した. この時, Decoder の LSTM と <BOS> および <EOS> の分散表現は MNLM に倣い全言語で共有する.

提案手法のハイパーパラメータは付録の表 A·1 でまとめた.

3.4 モデル選択

教師ありの手法では, モデル選択に開発データ 1000 文を用いた. RCSLS はエポック回数を 1 - 20 回, S-ID は $10K$ 回 (K=1,2,...30) で独立に学習し, 文のアラインメント精度 (i.e., 開発データのソース言語の各文に最も類似度 (CSLS) が高い文をターゲットの開発データ全文から抽出し, 対訳文かどうかの正解率を測った. 文の表現は単語の分散表現の和で表した.) が最も高かったモデルを評価に用いた. なお, XLM は 50 エポック学習して開発データの MLM perplexity が最小となったモデル, 提案手法は 20 エポック学習して生成の perplexity が最小となったモデルを用いた.

3.5 結果

表 1 に実験 1, 2, 3 の結果を示す. 提案手法 (OURS) が全ての言語ペアにおいて, 既存手法を大きく上回った. また, 提案手法で言語モデルを学習した場合 (OURS (LM)) も, 実験 2, 3 の全言語ペアで先行研究の MNLM を上回った. ベースライン同士を比較すると, マッピングの手法 (VecMap, Procrustes, RCSLS) は同時学習と比べると全体的に精度が低い結果となった. これは, 単語分散表現を学習する文数が少ないため, マップすることが困難であったためと考えれる. この結果は先行研究 [6] の結果と矛盾しない. また, XLM の精度も総じて低くなり, 特に実験 3 の日本語-英語ペアで精度がとりわけ低くなっている. 考えられる 1 つの原因として, これらの言語間では同じスペルの単語が非常に少ないため, 全言語で共通の語彙 (Shared Vocabulary) を学習することが困難だということがある. 従って, 文字や語順が異なる言語間で, かつ小規模のデータ (各言語 22,456 文) からマルチリンガルな単語表現を XLM で学習するのは困難だと示された.

スペイン語-英語の精度を比べると , S-ID や XLM では 2 言語で学習した時 (es-en (bi)) の方が 4 言語で学習した時 (es-en) よりも精度が低くなっている. 一方, 提案手法ではどちらの場合も高い精度となっている. 従って, 対訳コーパスの言語数やデータが少ない場合でも本手法は有効的であると明らかとなった. また, 実験 1, 2 の方が実験 3 よりも全ての手法で精度が全体的に高くなった. これは, ヨーロッパ言語の語順や語彙の近さが要因だと推測される.

4. 分析

4.1 Ablation Studies

提案手法で用いた各手法がどれほど有効であったかを確かめるために, 実験 2, 3 で手法の 1 個抜き実験 (Ablation Studies) を行なって精度を比較した (実験 2 は全言語ペアの平均値を比較した). 表 2 がその結果である. 実験 3 では, 特に入力文の生成 (自己符号化) がマルチリンガルな表現獲得において極めて重要な役割を果たしていることがわかる. これは, 実験 3 では日英または英日の 1 対 1 翻訳のみしか学習しないため, 両言語を同じ表現から同時に生成する制約が必要だからと推測される. 一方, 実験 2 では 1 言語の入力から多言語を生成するため, ある程度同じ空間に埋め込まれることが保証され, 自己符号化をしなくてもマルチリンガルの表現を獲得できる結果となった. また, Weight Tying が両実験において非常に有効的であることがわかった. 一般的に, Weight Tying が翻訳生成の精度を大幅に改善することはないため [14], この結果は興味深い. Weight Tying が本提案手法の精度を大きく向上させたのは, 入力と出力の単語分散表現の関係が注意機構等を通して効率的に学習出来たからだと考えられる. また, 単語分散表現の Dropout や逆向き生成もそれぞれ効果的である

実験 / 言語ペア	実験 2	実験 3
提案手法	65.5	47.2
w/o 逆向き生成	63.9	46.2
w/o Dropout	63.0	45.2
w/o Weight Tying	55.6	38.4
w/o 自己符号化	64.1	0.4

表 2: 実験 2, 3 における Ablation Studies. 実験 2 の値は全言語ペアの精度の平均である.

ことも明らかとなった.

5. 関連研究

5.1 マルチリンガル単語分散表現

マルチリンガル単語分散表現の既存手法には, 大きく分けて 2 つのアプローチが存在する. 一つは, word2vec[5] や FastText[25] 等の手法で事前学習された各言語の分散表現空間を共通空間に写像する方法, もう一つは対訳コーパス等を用いてマルチリンガルな分散表現を同時学習する手法である. 本稿では前者を「**A. マッピング**」, 後者を「**B. 同時学習**」と呼ぶ. なお, 本研究は後者の「同時学習」の手法に属する. 以下, それぞれの関連研究を簡潔に説明する.

A. マッピング

マッピングの手法の多くは, 対訳辞書を用いて空間をマップする行列 W 学習する [28], [29]. 対訳辞書に存在するソース言語の単語分散表現 X_i と対応するターゲット言語の単語分散表現 Y_i が近くなるように線形写像 W を学習する.

$$W^* = \arg\min_W \sum_i ||X_i W - Y_i||^2 \qquad (13)$$

この W に直交行列の制約を課すことでマッピングの精度が更に向上すると知られている [29]. そのような制約下では Procrustes Solution [21] と呼ばる解が存在し, YX^T を特異値分解 (SVD) することで以下のように解が得られる.

$$\begin{aligned} W^* &= \arg\min_W \sum_i ||X_i W - Y_i||^2 = UV^{\mathrm{T}} \\ s.t. \quad & U\Sigma V^{\mathrm{T}} = \mathrm{SVD}(YX^{\mathrm{T}}) \end{aligned} \qquad (14)$$

ここで, Σ は対角行列である.

近年では, 対訳データを一切用いずに写像 W を学習する教師なしの手法も数多く提案されている [19], [20], [30]. 2 言語の分散表現の分布間の距離を最小化する手法 [31], [32] や, 敵対的学習を用いて元言語が識別できなくなるようにマッピングする手法 [19], [30] が存在する. また, 擬似対訳辞書の生成と線形写像の学習を繰り返す自己学習を用いた手法 [20] も存在する.

マッピング手法に共通する問題は, 対応づける 2 言語の単語分散空間で直交写像が存在する, 言い換えると 2 言語の分散表現空間のグラフがおよそ同型でということを仮定する点である. 一般にその仮定は成り立たず, 特に言語距離が遠い場合や低資源の状況下では学習が困難であると知られている [6], [7].

B. 同時学習

同時学習の手法は, 主に単語や文, 文書のアラインメント情報を用いて分散表現を学習する. [8] は対訳コーパスでアラインされた文のセットに ID を割り当て, 文中の単語から ID を Skip Gram で予測する手法である. これにより, 同じ ID を持つ文（i.e., 対訳文）の単語分散表現が近くなりマルチリンガルな単語分散表現を獲得できる. Bivec [11] は更に対訳文間の単語アラインメントの情報を用いる手法で, 各単語の周辺単語と対訳文でアラインされた単語の周辺単語を同時に予測する学習を Skip Gram で行う. また, 対訳コーパスから対応する単語またはフレーズを抽出し, アラインされた単語やフレーズの分散表現が近くなるように学習する手法もいくつか存在する [9], [10]. 対訳コーパスの代わりに対訳辞書を用いる手法もいくつか存在し, 例えば単言語コーパスの一部の単語を他言語の単語に辞書で置換し, word2vec などで同時学習を行う手法も存在する [33].

またマッピングと同様に, 対訳データを用いずに学習する教師なしの手法が近年提案されている [6], [23], [34]. XLM [23] は大規模データで学習する BERT [26] のような Masked 言語モデル (MLM) を多言語で共有して学習することにより, マルチリンガルな文およびサブワード表現を獲得する. XLM は対訳データを用いた教師ありの手法にも応用可能であり, その場合は Translation 言語モデル (TML) を MLM と同時に学習する. また, 先行研究 [6] では, LSTM の言語モデルを語順の近い言語間でシェアすることで, 低資源下においてもマルチリンガルな単語分散表現を獲得できることを示した. また, 教師なし翻訳モデル [35] で生成した擬似対訳コーパスに対して, 先述の文 ID に基づく手法 [8] や Bivec [11] を学習することで, 教師なしマッピングの精度を上回ることも知られている [34].

5.2 多言語機械翻訳モデル

提案手法は多言語機械翻訳モデルを学習することで単語分散表現を獲得する. 多言語機械翻訳とは, 一つのモデルで複数言語の翻訳文を生成する手法である. 各言語で独立に学習する場合と比べて少ないパラメータで学習することができ, また学習データが存在しない言語間で翻訳を行う事も可能となる (zero-shot 翻訳) [36]. また, 本研究のように多言語機械翻訳モデルを他のマルチリンガルタスクに応用する研究も存在する. 例えば, 先行研究 [37] では, 多言語機械翻訳モデルの Encoder を用いてマルチリンガルな文表現を獲得することを試みた. 対訳文コーパスの文集合から対訳文のペアをアラインするタスクで評価を行い, Encoder の文表現が全言語共通の空間に埋め込まれていることを示した. しかし, この研究ではモデルをサブワードを用いて

学習しており, 単語分散表現に着目したモデル設計や評価は行われていない.

6. まとめ

本研究は, Encoder-Decoder モデルと対訳コーパスを用いてマルチリンガル単語分散表現を学習する新たな手法を提案した. 本研究が用いた Encoder-Decoder モデルは, Decoder の LSTM と単語分散表現のパラメータを言語毎に学習し, それ以外の全てのパラメータを全言語で共有した. 評価には対訳辞書生成 (Billingual Lexicon Induction) の実験を 3 種類の言語セットで行い, 全ての条件で提案手法が既存手法を大幅に上回ることを示した. 今後は, 提案手法を中規模から大規模の対訳コーパス (100 万文超) で学習した時の精度や, 単言語コーパスを利用した半教師あり学習の有効性, さらにはマイナー言語への転移学習の可能性について調べていきたい. また, 本研究による単語アラインメントの精度を IBM models [38] のような伝統的な手法と比較することも検討している.

参考文献

[1] Lample, G., Conneau, A., Denoyer, L. and Ranzato, M.: Unsupervised Machine Translation Using Monolingual Corpora Only, *International Conference on Learning Representations* (2018).

[2] Adams, O., Makarucha, A., Neubig, G., Bird, S. and Cohn, T.: Cross-Lingual Word Embeddings for Low-Resource Language Modeling, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, Association for Computational Linguistics, pp. 937–947 (online), available from ⟨http://aclweb.org/anthology/E17-1088⟩ (2017).

[3] Xie, J., Yang, Z., Neubig, G., Smith, N. A. and Carbonell, J.: Neural Cross-Lingual Named Entity Recognition with Minimal Resources, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, Association for Computational Linguistics, pp. 369–379 (online), DOI: 10.18653/v1/D18-1034 (2018).

[4] Xiao, M. and Guo, Y.: Distributed Word Representation Learning for Cross-Lingual Dependency Parsing, *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, Association for Computational Linguistics, pp. 119–129 (online), DOI: 10.3115/v1/W14-1613 (2014).

[5] Mikolov, T., Chen, K., Corrado, G. and Dean, J.: Efficient Estimation of Word Representations in Vector Space, *International Conference on Learning Representations (Workshop)* (2013).

[6] Wada, T., Iwata, T. and Matsumoto, Y.: Unsupervised Multilingual Word Embedding with Limited Resources using Neural Language Models, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, Association for Computational Linguistics, pp. 3113–3124 (online), available from ⟨https://www.aclweb.org/anthology/P19-1300⟩ (2019).

[7] Søgaard, A., Ruder, S. and Vulić, I.: On the Limitations of Unsupervised Bilingual Dictionary Induction, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, pp. 778–788 (online), available from ⟨http://aclweb.org/anthology/P18-1072⟩ (2018).

[8] Levy, O., Søgaard, A. and Goldberg, Y.: A Strong Baseline for Learning Cross-Lingual Word Embeddings from Sentence Alignments, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, Valencia, Spain, Association for Computational Linguistics, pp. 765–774 (online), available from ⟨https://www.aclweb.org/anthology/E17-1072⟩ (2017).

[9] Dufter, P., Zhao, M., Schmitt, M., Fraser, A. and Schütze, H.: Embedding Learning Through Multilingual Concept Induction, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia, Association for Computational Linguistics, pp. 1520–1530 (online), DOI: 10.18653/v1/P18-1141 (2018).

[10] Dufter, P. and Schütze, H.: A Stronger Baseline for Multilingual Word Embeddings, *CoRR*, Vol. abs/1811.00586 (online), available from ⟨http://arxiv.org/abs/1811.00586⟩ (2018).

[11] Luong, M.-T., Pham, H. and Manning, C. D.: Bilingual Word Representations with Monolingual Quality in Mind, *NAACL Workshop on Vector Space Modeling for NLP*, Denver, United States (2015).

[12] Takeno, S., Nagata, M. and Yamamoto, K.: Controlling Target Features in Neural Machine Translation via Prefix Constraints, *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*, Taipei, Taiwan, Asian Federation of Natural Language Processing, pp. 55–63 (online), available from ⟨https://www.aclweb.org/anthology/W17-5702⟩ (2017).

[13] Inan, H., Khosravi, K. and Socher, R.: Tying Word Vectors and Word Classifiers: A Loss Framework for Language Modeling, *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, (online), available from ⟨https://openreview.net/forum?id=r1aPbsFle⟩ (2017).

[14] Press, O. and Wolf, L.: Using the Output Embedding to Improve Language Models, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, Valencia, Spain, Association for Computational Linguistics, pp. 157–163 (online), available from ⟨https://www.aclweb.org/anthology/E17-2025⟩ (2017).

[15] Luong, T., Pham, H. and Manning, C. D.: Effective Approaches to Attention-based Neural Machine Translation, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, Association for Computational Linguistics, pp. 1412–1421 (online), DOI: 10.18653/v1/D15-1166 (2015).

[16] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R.: Dropout: A Simple Way to Prevent Neural Networks from Overfitting, *Journal of Machine Learning Research*, Vol. 15, pp. 1929–1958 (online), available from ⟨http://jmlr.org/papers/v15/srivastava14a.html⟩ (2014).

[17] Christodouloupoulos, C. and Steedman, M.: A massively parallel corpus: the Bible in 100 languages, *Language Resources and Evaluation*, Vol. 49, No. 2, pp. 375–395 (online), available from

ⓒ 2019 Information Processing Society of Japan

⟨https://doi.org/10.1007/s10579-014-9287-y⟩ (2015).

[18] Straka, M., Hajič, J. and Straková, J.: UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Portorož, Slovenia, European Language Resources Association (ELRA), pp. 4290–4297 (online), available from ⟨https://www.aclweb.org/anthology/L16-1680⟩ (2016).

[19] Conneau, A., Lample, G., Ranzato, M., Denoyer, L. and Jégou, H.: Word translation without parallel data, *International Conference on Learning Representations* (2018).

[20] Artetxe, M., Labaka, G. and Agirre, E.: A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, pp. 789–798 (online), available from ⟨http://aclweb.org/anthology/P18-1073⟩ (2018).

[21] Schönemann, P. H.: A generalized solution of the orthogonal procrustes problem, *Psychometrika*, Vol. 31, No. 1, pp. 1–10 (1966).

[22] Joulin, A., Bojanowski, P., Mikolov, T., Jégou, H. and Grave, E.: Loss in Translation: Learning Bilingual Word Mapping with a Retrieval Criterion, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, Association for Computational Linguistics, pp. 2979–2984 (online), DOI: 10.18653/v1/D18-1330 (2018).

[23] Lample, G. and Conneau, A.: Cross-lingual Language Model Pretraining, *arXiv preprint arXiv:1901.07291* (2019).

[24] Glavaš, G., Litschko, R., Ruder, S. and Vulić, I.: How to (Properly) Evaluate Cross-Lingual Word Embeddings: On Strong Baselines, Comparative Analyses, and Some Misconceptions, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, Association for Computational Linguistics, pp. 710–721 (online), DOI: 10.18653/v1/P19-1070 (2019).

[25] Bojanowski, P., Grave, E., Joulin, A. and Mikolov, T.: Enriching Word Vectors with Subword Information, *Transactions of the Association for Computational Linguistics*, Vol. 5, pp. 135–146 (online), available from ⟨http://aclweb.org/anthology/Q17-1010⟩ (2017).

[26] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota, Association for Computational Linguistics, pp. 4171–4186 (online), DOI: 10.18653/v1/N19-1423 (2019).

[27] Sennrich, R., Haddow, B. and Birch, A.: Neural Machine Translation of Rare Words with Subword Units, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany, Association for Computational Linguistics, pp. 1715–1725 (online), DOI: 10.18653/v1/P16-1162 (2016).

[28] Mikolov, T., Le, Q. V. and Sutskever, I.: Exploiting Similarities among Languages for Machine Translation., *CoRR*, Vol. abs/1309.4168 (online), available from ⟨https://arxiv.org/pdf/1309.4168.pdf⟩ (2013).

[29] Xing, C., Wang, D., Liu, C. and Lin, Y.: Normalized Word Embedding and Orthogonal Transform for Bilingual Word Translation, *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, pp. 1006–1011 (online), DOI: 10.3115/v1/N15-1104 (2015).

[30] Zhang, M., Liu, Y., Luan, H. and Sun, M.: Adversarial Training for Unsupervised Bilingual Lexicon Induction, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, pp. 1959–1970 (online), DOI: 10.18653/v1/P17-1179 (2017).

[31] Zhang, M., Liu, Y., Luan, H. and Sun, M.: Earth Mover's Distance Minimization for Unsupervised Bilingual Lexicon Induction, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, pp. 1934–1945 (online), available from ⟨http://aclweb.org/anthology/D17-1207⟩ (2017).

[32] Xu, R., Yang, Y., Otani, N. and Wu, Y.: Unsupervised Cross-lingual Transfer of Word Embedding Spaces, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, pp. 2465–2474 (online), available from ⟨http://aclweb.org/anthology/D18-1268⟩ (2018).

[33] Gouws, S. and Søgaard, A.: Simple task-specific bilingual word embeddings, *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Denver, Colorado, Association for Computational Linguistics, pp. 1386–1390 (online), DOI: 10.3115/v1/N15-1157 (2015).

[34] Marie, B. and Fujita, A.: Unsupervised Joint Training of Bilingual Word Embeddings, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, Association for Computational Linguistics, pp. 3224–3230 (online), DOI: 10.18653/v1/P19-1312 (2019).

[35] Lample, G., Ott, M., Conneau, A., Denoyer, L. and Ranzato, M.: Phrase-Based & Neural Unsupervised Machine Translation, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2018).

[36] Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M. and Dean, J.: Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation, *Transactions of the Association for Computational Linguistics*, Vol. 5, pp. 339–351 (online), available from ⟨http://aclweb.org/anthology/Q17-1024⟩ (2017).

[37] Schwenk, H. and Douze, M.: Learning Joint Multilingual Sentence Representations with Neural Machine Translation, *Proceedings of the 2nd Workshop on Representation Learning for NLP*, Vancouver, Canada, Association for Computational Linguistics, pp. 157–167 (online), DOI: 10.18653/v1/W17-2619 (2017).

[38] Brown, P. F., Della Pietra, S. A., Della Pietra, V. J. and Mercer, R. L.: The Mathematics of Statistical Machine Translation: Parameter Estimation, *Computational Linguistics*, Vol. 19, No. 2, pp. 263–311 (online), available from ⟨https://www.aclweb.org/anthology/J93-2003⟩ (1993).

ⓒ 2019 Information Processing Society of Japan

[39] Pascanu, R., Mikolov, T. and Bengio, Y.: On the difficulty of training recurrent neural networks, *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, pp. 1310–1318 (online), available from ⟨http://jmlr.org/proceedings/papers/v28/pascanu13.html⟩ (2013).

付　　録

A.1 ハイパーパラメータ

ハイパーパラメータ	教師あり	教師なし
Encoder LSTM 層数	1	0
Decoder LSTM 層数	2	
LSTM 隠れ層 次元数	500	
単語分散表現 (E) 次元数	500	
最適化手法	Adam	SGD
エポック数	20	50
ミニバッチサイズ	64	
学習率	0.001	1.0
Gradient Clipping [39]	5.0	
Word Dropout Rate	0.5	

表 A·1: 提案手法のハイパーパラメータ一覧

ⓒ 2019 Information Processing Society of Japan