双方向翻訳のための中間表現制約を用いたニューラル機械翻訳

小林 尚輝[1]，　田村 晃裕[2]，　二宮 崇[2]，　高村 大也[3,4]，　奥村 学[3]
[1]愛媛大学 情報工学科，[2]愛媛大学 大学院理工学研究科 電子情報工学専攻，
[3]東京工業大学 科学技術創成研究院，[4]産業技術総合研究所
{kobayasi@.ai.,tamura@,ninomiya@}cs.ehime-u.ac.jp,
{takamura,oku}@pi.titech.ac.jp

1　はじめに

近年，自然言語処理の多くのタスクにおいて，ニューラルネットワークを用いた手法が盛んに研究され，特に系列から系列への変換を学習するエンコーダ･デコーダモデル [2, 6] は従来手法を上回る高い性能を実現することから注目されている．エンコーダ・デコーダモデルは，エンコーダ及びデコーダと呼ばれる 2 つの再帰型ニューラルネットワークから構成されており，エンコーダは入力された系列を中間表現へと変換し，デコーダは中間表現から出力となる系列を生成することで，系列から系列への変換を行う．多くのニューラル機械翻訳は，エンコーダ・デコーダモデルによって実現されており，対訳関係にある文対をそれぞれ入力系列 (原言語文) と出力系列 (目的言語文) とすることで，原言語から目的言語への翻訳モデルを学習する．さらに，ニューラル機械翻訳にアテンション機構を導入することで，デコーダはエンコーダの隠れ状態の履歴を参照しながらデコードすることができ，入力系列が長い場合においても良い出力系列を生成できる．

アテンションに基づくニューラル機械翻訳は，従来の統計的機械翻訳よりも高い翻訳精度を実現するが，その多くの手法は，対訳文書のいずれか一方の言語を原言語とし，もう片方の言語を目的言語とすることで翻訳モデルを学習するため，一度に学習するモデルは単方向の翻訳モデルだけであった．言語間双方向の翻訳モデルを同時に学習することができれば，翻訳方向が異なるエンコーダ-デコーダのモデル共有化や，翻訳方向間で翻訳モデルの整合性がとれ，より性能の高い翻訳モデルを学習できることが期待される．

本論文は，中間表現制約を用いた双方向翻訳のためのエンコーダ・デコーダモデル同時学習方法を提案する．エンコーダとデコーダはそれぞれの役割に応じて異なるネットワーク構成をしているが，どちらも単語を入力とする再帰型ニューラルネットワークを基本構成要素としている．そのため，本研究では翻訳方向が異なるエンコーダとデコーダ間での重みの共有化を行い，エンコーダのユニットはデコーダの一部としても学習を行い，デコーダのユニットはエンコーダの一部としての学習も行う．しかし，翻訳方向が異なるエンコーダ-デコーダ間で重みを共有化した場合，エンコーダとデコーダが学習する中間表現の埋め込み空間が異なっていると，正しくエンコード/デコードすることができないという問題が生じる．本研究は，その問題を解消するため，それぞれの言語のエンコーダの生成する中間表現が同一のベクトルになる制約を導入する．

日英の対訳コーパスを用いて実験を行い，日英，英日における双方向の機械翻訳モデルの同時学習と中間表現制約により BLEU スコアが 0.87 ポイント改善したことを確認した．

関連研究として，双方向翻訳のためのエンコーダ・デコーダモデルにおいてアテンションに基づく制約を与える手法 [1] があげられる．彼らは，アテンションに基づく単語間アライメントが双方向で一致するように学習しているが，本研究は中間表現が一致するように学習する点で異なっている．

2　ニューラル機械翻訳

2.1　エンコーダ・デコーダモデル

エンコーダ・デコーダモデル [2, 6] は，エンコーダを用いて入力文を固定長のベクトル表現へと変換し，デコーダを用いて逐次的に単語の予測を行うモデルである．本研究では，エンコーダとデコーダにそれぞれ LSTM[3] を用い，それぞれ $LSTM_{enc}$, $LSTM_{dec}$ と表す．LSTM は前時刻の隠れ状態 h_{t-1} と入力 x_t から隠れ状態 h_t を生成する再帰型ニューラルネットワー

クであり，隠れ状態 h_t は次式により定義される:

$$
\begin{aligned}
i_t &= \sigma(W^{(i)}x_t + U^{(i)}h_{t-1} + b^{(i)}),\\
f_t &= \sigma(W^{(f)}x_t + U^{(f)}h_{t-1} + b^{(f)}),\\
o_t &= \sigma(W^{(o)}x_t + U^{(o)}h_{t-1} + b^{(o)}),\\
\bar{h}_t &= \tanh(W^{(\bar{h})}x_t + U^{(\bar{h})}h_{t-1} + b^{(\bar{h})}),\\
c_t &= i_t \odot \bar{h}_t + f_t \odot c_{t-1},\\
h_t &= o_t \odot \tanh(c_t).
\end{aligned}
$$

ただし，$W^{(i)}, W^{(f)}, W^{(o)}, W^{(\bar{h})} \in \mathcal{R}^{H\times I}$, $U^{(i)}, U^{(f)}, U^{(o)}, U^{(\bar{h})} \in \mathcal{R}^{H\times H}$, $b^{(i)}, b^{(f)}, b^{(o)}, b^{(\bar{h})} \in \mathcal{R}^{H\times 1}$ は重みであり，I は入力 x_t の次元，H は隠れ状態の次元である．また，σ はロジスティックシグモイド関数，$\odot$ はベクトルの要素積を表す．

エンコーダ・デコーダモデルは，入力系列 $\boldsymbol{x} = (x_1, \ldots, x_n)$ が与えられたとき，出力系列 $\boldsymbol{y} = (y_1, \ldots, y_m)$ の対数尤度を最大化することを目的とし，次式を用いて定式化される:

$$\log p(\boldsymbol{y}|\boldsymbol{x}) = \sum_{j=1}^{m} \log p(y_j|y_{<j}, \boldsymbol{x}). \tag{1}$$

エンコーダは入力系列 $\boldsymbol{x}$ に対し，次式で計算される中間表現 $\boldsymbol{c} = \{\bar{h}_1, \ldots, \bar{h}_n\}$ を生成する:

$$\bar{h}_j = LSTM_{enc}(\bar{h}_{j-1}, x_j). \tag{2}$$

デコーダの隠れ状態は初期状態を $h_0 = \bar{h}_n$ とし，時刻 j におけるデコーダの隠れ状態 h_jは 直前の出力 y_{j-1} と隠れ状態 h_{j-1} から次式により計算される:

$$h_j = LSTM_{dec}(h_{j-1}, y_{j-1}). \tag{3}$$

単語の出力確率は，デコーダの隠れ状態 h_j を用いて次式で計算される:

$$p(y_j|y_{<j}, \boldsymbol{x}) = softmax(W_o h_j). \tag{4}$$

ただし，$W_o \in \mathcal{R}^{V\times H}$ は重み行列であり，V, H はそれぞれ辞書の大きさ，隠れ状態の次元を表す．

エンコーダ・デコーダモデルの目的関数は次式で表される:

$$J = \sum_{(x,y)\in D} -\log p(y|x). \tag{5}$$

ここで，D はデータセット全体を表す．

2.2 アテンションに基づくニューラル機械翻訳

アテンション機構 [2, 6] は，各時刻におけるデコーダの単語予測に用いる文脈ベクトルをエンコーダ側の隠れ状態の履歴から獲得するための手法である．

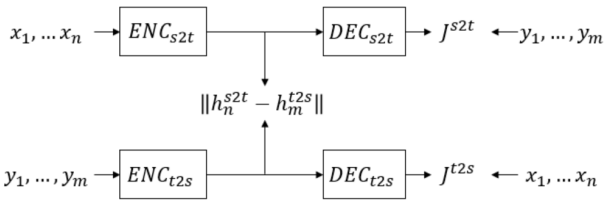

図 1: 提案手法の学習

文脈ベクトル c_t はエンコーダの全隠れ状態 $S = \{\bar{h}_1, \ldots, \bar{h}_n\}$ の加重平均として次式により計算される:

$$c_t = \sum_{s\in S} \alpha_t(s)\bar{h}_s. \tag{6}$$

ここで，重み $\alpha_t(s)$ はデコーダの隠れ状態 h_t に対するエンコーダの隠れ状態 $\bar{h_s}$ の重要度を表し，次式で計算される:

$$\alpha_t(s) = \frac{exp(h_t \cdot \bar{h}_s)}{\sum_{s'\in S} \exp(h_t \cdot \bar{h}_{s'})}. \tag{7}$$

アテンション機構付きエンコーダ・デコーダモデルでは，次式で計算される $\hat{h}$ を，式 (4) における h_j として計算する:

$$\hat{h} = \tanh(W_c[c_t; h_t]). \tag{8}$$

ここで,$W_c \in \mathcal{R}^{H\times 2H}$ は重み行列である．

3 提案手法

本節は，提案手法である中間表現制約を導入した双方向翻訳のためのエンコーダ・デコーダモデルについて説明する．まず，双方向翻訳モデルの重みの共有について説明し，次に中間表現制約と目的関数を説明する．

3.1 重み共有

双方向翻訳のためのエンコーダ・デコーダモデルにおける重み共有について説明する．図 1 は提案モデルの全体像を示す．単一方向への翻訳モデルはエンコーダとデコーダの 2 つの LSTM から構成されるが，提案モデルである双方向翻訳のためのエンコーダ･デコーダモデルでは，図 1 に示されるように，計 4 つの LSTM を学習する必要がある．原言語から目的言語へと翻訳するエンコーダ，デコーダを ENC_{s2t}，DEC_{s2t} とし，目的言語から原言語へと翻訳するエンコーダ，デコーダを ENC_{t2s}，DEC_{t2s} と表す．各 LSTM に入

力される言語は，ENC_{s2t}，DEC_{t2s} が原言語となり，ENC_{t2s}，DEC_{s2t} が目的言語となる．扱う言語が同じ LSTM の重みを共有することにより，エンコーダの LSTM はデコーダの一部としても学習を行い，デコーダの LSTM はエンコーダの一部としての学習も行われる．

3.2 中間表現制約を用いた目的関数

双方向翻訳のためのエンコーダ・デコーダモデルにおいて中間表現制約を導入した目的関数について説明する．対訳関係にある $\boldsymbol{x} = (x_1, \ldots, x_n) \in L_1$, $\boldsymbol{y} = (y_1, \ldots, y_m) \in L_2$ に対して，式 (9)，式 (10) よりエンコーダの隠れ状態を求める:

$$h_j^{s2t} = ENC_{s2t}(h_{j-1}^{s2t}, x_j), \tag{9}$$

$$h_j^{t2s} = ENC_{t2s}(h_{j-1}^{t2s}, y_j). \tag{10}$$

本手法では，それぞれのエンコーダの最終状態 h_n^{s2t}, h_m^{t2s} の二乗誤差を最小化することで，中間表現が同一のベクトルになるよう学習する．式 (5) により，言語 L_1 から言語 L_2 への翻訳における目的関数を J^{s2t}，言語 L_2 から言語 L_1 への翻訳における目的関数を J^{t2s} とすると，提案手法の目的関数は次式により与えられる:

$$L = J^{s2t} + J^{t2s} + ||h_n^{s2t} - h_m^{t2s}||^2. \tag{11}$$

4 実験

4.1 コーパス

本実験では，日英の対訳コーパスである，Asian Scientific Paper Excerpt Corpus (ASPEC)[7] を用いた．英語の単語分割は moses decoder [1] を用い，日本語の単語分割は KyTea[2]を用いた．全てのアルファベットは小文字化した．

データセットの整形は橋本ら [4] に習い，学習データ (train-1.txt) から抜き出した単語数が 50 以下からなる文対のみを使用し，上位 2 万文対と 10 万文対の 2 通りを用いて学習を行った．語彙は単語の出現頻度が 2 回以上のものを使用し，語彙に含まれない単語は <UNK> トークンに置き換えた．学習したモデルの評価には開発データ (dev.txt: 1790 文対) を用いた．

[1]https://github.com/moses-smt/mosesdecoder/blob/master/scripts/tokenizer/tokenizer.perl

[2]http://www.phontron.com/kytea/index-ja.html

表 1: 実験結果

手法	BLEU			
	英 → 日		日 → 英	
	20k	100k	20k	100k
ベースラインモデル	13.57	23.96	9.30	17.80
重み共有モデル	13.04	24.06	8.87	17.65
提案手法	14.34	24.83	9.98	18.02

4.2 評価手法

翻訳精度の評価手法として BLEU[8] を用いた．開発データに対して最も性能の良いモデルを選択し，テストデータ (test.txt: 1812 文対) を用い手法の評価を行った．

4.3 モデルのパラメータ

単語埋め込み層，隠れ層の次元はともに 256 とし，ミニバッチサイズは 2 万文，10 万文でそれぞれ 256, 128 とした．最適化手法に Adam[5] を用い，$\alpha = 0.003$ とした．正則化は L-2 正則化と dropout を用い，それぞれの係数を $1.0E-6$，0.2 とした．勾配クリッピングの値は 3.0 とした．

4.4 結果

実験結果を表 1 に示す．ベースラインモデルは日英，英日の翻訳モデルを別々に学習した場合である．ベースラインの構成はエンコーダが 2 stacked Bidirectional LSTM，デコーダが 2 stacked LSTM であり，アテンション付きのモデルになっている．重み共有モデルは，ベースラインの日英，英日モデルを重み共有し同時学習したモデルである．提案手法は，重み共有モデルに中間状態に関する制約を付与したモデルである．日英翻訳，英日翻訳をそれぞれ 2 万文，10 万文で学習させた結果，中間表現制約を付与した提案手法はベースラインモデルを最大 0.87 ポイント上回った．

5 考察

重み共有モデルとベースラインモデルの BLEU スコアには差がみられなかった．言い換えれば，LSTM の重みを共有しても精度が低下しないと考えられる．

表 2: 中間表現間の距離

手法	距離
重み共有モデル	39.63
提案手法	0.71

提案手法に関しては，中間表現制約により BLEU スコアが向上した考えられる．

重み共有モデルと提案手法について英日，日英の中間表現間の距離をテストデータを対象に計測した結果を表 2 に示す．重み共有モデルに比べ，中間表現制約を加えた提案手法は中間表現間の距離が大幅に小さくなっていることが確認できる．

以上の結果から，中間表現制約によって対訳関係にある文に対して類似した中間表現が得られ，BLEU スコアの向上がみられることが分かった．異なる言語に対して類似した中間表現を得られることから，ピボット翻訳への応用が期待できる．具体的には，英語をピボットとした日仏翻訳を行う際に，日英のエンコーダにより得られた中間表現を英仏のデコーダの初期状態として入力することで，英文を介さずに直接日仏の翻訳が可能となり情報の欠落を抑える効果が期待される．

6 おわりに

本研究では，エンコーダ・デコーダモデルによって生成される中間表現に注目し，それぞれの言語で生成される中間表現が同一となるように制約をかける手法を提案した．そして，日英および英日翻訳においてともに提案手法の有効性を確認した．三か国語以上を用いた実験および大規模なデータを用いた実験については今後の課題としたい．

7 謝辞

本研究は JSPS 科研費 25280084 の助成を受けたものである．ここに謝意を表する．

参考文献

[1] Yong Cheng, Shiqi Shen, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. Agreement-based joint training for bidirectional attention-based neural machine translation. In *Proceedings of the IJCAI*, 2016.

[2] Bahdanau Dzmitry, Cho KyungHyun, and Bengio Yoshua. Neural machine translation by jointly learning to align and translate. In *Proceedings of the ICLR*, 2015.

[3] Felix A. Gers, Jürgen Schmidhuber, and Fred A. Cummins. Learning to forget: Continual prediction with lstm. *Neural computation*, 2000.

[4] Kazuma Hashimoto and Yoshimasa Tsuruoka. Neural Machine Translation with Source-Side Latent Graph Parsing. In *Proceedings of the EMNLP*, 2017.

[5] Diederik Kingsma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the ICLR*, 2014.

[6] Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the EMNLP*, 2015.

[7] Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. Aspec: Asian scientific paper excerpt corpus. In *Proceedings of the LREC*, 2016.

[8] Kishore Papineni, Salam Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the ACL*, 2002.