文レベルの中間表現とのアテンションを用いたニューラル機械翻訳

福原 健太 [1]　　田村 晃裕 [2]　　二宮 崇 [2]

[1] 愛媛大学 工学部情報工学科

[2] 愛媛大学 大学院理工学研究科 電子情報工学専攻

{hukuhara@ai., tamura@, ninomiya@}cs.ehime-u.ac.jp

1　はじめに

機械翻訳は，これまでルールベース機械翻訳や統計的機械翻訳などの様々な手法が提案されてきたが，現在は，ニューラルネットワークを用いた機械翻訳（以下ニューラル機械翻訳) が主流となっている．ニューラル機械翻訳として Recurrent Neural Network（RNN）を用いたエンコーダー・デコーダーモデル [1] が初期から広く使われている．エンコーダー・デコーダーモデルは入力系列（原言語文の単語系列）を中間表現に変換するエンコーダーと，エンコーダーで変換された中間表現をもとに出力系列（目的言語文の単語系列）を生成するデコーダーで構成されている．エンコーダー・デコーダーモデルは，デコーダーで出力単語を決める際に，エンコーダーの各隠れ状態を参照して入力系列の中で注目する部分を捉えるアテンション機構を導入することで翻訳性能を改善している [2].

エンコーダー・デコーダーモデルは機械翻訳以外の様々なタスクにおいても利用されており，テキストから画像を生成する画像生成タスクにおいても成果をあげている．画像生成タスクにおけるエンコーダー・デコーダーモデルの一つとして MirrorGAN[3] が提案されている．MirrorGAN は，入力テキストを RNN によりエンコードし，エンコード結果とアテンションに基づき敵対的生成ネットワークにより画像を生成する．その際，各入力単語の中間表現とのアテンションに加えて，文レベルの中間表現とのアテンションを加えることで画像生成性能を向上させている．

一方で，機械翻訳タスクのアテンション機構付き NMT においては，通常，アテンション機構では各入力単語の中間表現（エンコーダーの各隠れ状態）とのアテンションしか考慮されない．そこで，本論文では，アテンション機構付き NMT モデル [2] において，エンコーダーが変換した文レベルの中間表現（エンコーダーの最終隠れ状態）とのアテンションも用いるモデルを提案する．ASPEC データを用いた日英翻訳の評価実験を通じて，文レベルの中間表現とのアテンションを考慮することで，BLEU が 0.72 ポイント向上することを確認した．

2　従来の NMT モデル

2.1　RNN に基づくエンコーダー・デコーダー NMT

RNN に基づくエンコーダー・デコーダー NMT モデル [1] は RNN エンコーダーで入力系列 $x = (x_1, x_2, ..., x_n)$ を固定長ベクトルの中間表現 $c = (\bar{h}_1, \bar{h}_2, ..., \bar{h}_n)$ に変換し，RNN デコーダーで変換された中間表現から出力系列 $y = (y_1, y_2, ..., y_m)$ を逐次的に生成する．RNN としては GRU や LSTM などが使用されるが，本研究では LSTM を使用する．

エンコーダーでは，i 番目の隠れ状態 h_i は 1 ステップ前の隠れ状態 h_{i-1} と現ステップの入力 x_i からエンコーダー LSTM により，式 (1) のように算出される．

$$\bar{h}_i = LSTM_{enc}(\bar{h}_{i-1}, E_x(x_i)). \quad (1)$$

ここで，$LSTM_{enc}$ はエンコーダー側の LSTM であり，E_x は単語埋め込み層である．

デコーダーでは，初期の隠れ状態 h_0 をエンコーダーの最終隠れ状態 $\bar{h}_n$ として，j 番目の出力単語に対する確率分布を式 (2)，(3) のように算出する．具体的には，まず，j 番目の LSTM デコーダーの隠れ状態 h_j を 1 ステップ前の隠れ状態 h_{j-1} と 1 ステップ前の出力単語 y_{j-1} から算出する（式 (2)）．その後，算出した h_j を目的言語の語彙数次元のベクトルに線形変換し，ソフトマックス関数を適用することで，出力単語

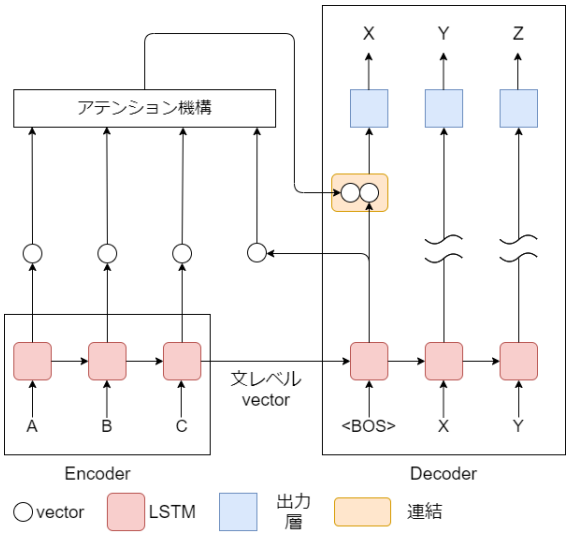

図 1: 従来のアテンション機構付きエンコーダー・デコーダーモデル

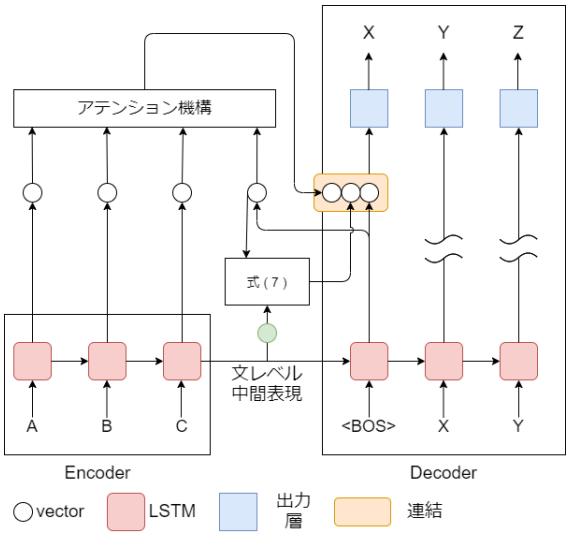

図 2: 提案モデル (提案手法 1)

に対する確率分布を獲得する（式 (3)）.

$$h_j = LSTM_{dec}(h_{j-1}, E_y(y_{j-1})). \tag{2}$$

$$p(y_j|y_{1:j-1}, h_n) = softmax(W_o h_j). \tag{3}$$

ここで，$LSTM_{dec}$ はデコーダー側の LSTM であり，E_y は単語埋め込み層である．また，$W_o \in R^{V \times H}$ は重み行列であり，V は目的言語の語彙数，H は LSTM の隠れ層の次元を表す．

2.2 アテンション機構

アテンション機構 [2] は，デコーダーの各ステップにおいて出力単語を生成する際に，LSTM エンコーダーの各隠れ状態を参照することで，長文においても翻訳精度の減少を抑える機構である．アテンション機構の概要図を図 1 に示す．

アテンション機構付きエンコーダー・デコーダーモデルでは，出力単語の確率分布を計算する際，式 (3) において，h_j の代りに以下の式 (4) で算出される $\hat{h}$ を用いる．

$$\hat{h} = tanh(W_c[c_t; h_j]) \tag{4}$$

ここで，$W_c \in R^{H \times 2H}$ は重み行列である．また，式 (4) における c_t は文脈ベクトルと呼ばれ，式 (5) のように，LSTM エンコーダーの各隠れ状態 $(\bar{h}_1, ..., \bar{h}_n)$ の加重平均（重みは α_t）である．

$$c_t = \sum_{i=1}^{n} \alpha_t(i)\bar{h}_i \tag{5}$$

重み $\alpha_t(i)$ は，LSTM デコーダーの隠れ状態 h_j と LSTM エンコーダーの隠れ状態 $\bar{h}_i$ との類似度を計算したもので式 (6) により計算される．

$$\alpha_t(i) = \frac{exp(h_j \cdot \bar{h}_i)}{\sum_{k=1}^{N} exp(h_j \cdot \bar{h}_k)} \tag{6}$$

3 提案モデル

本節では，アテンション機構の中で，単語レベルの中間表現とのアテンションとともに，文レベルの中間表現とのアテンションも用いるアテンション機構付き NMT モデルを提案する．提案モデルでは，エンコーダー LSTM の各隠れ状態 $(\bar{h}_1, \bar{h}_2, ..., \bar{h}_n)$ を単語レベルの中間表現，エンコーダーの最終隠れ状態 $\bar{h}_n$ を文レベルの中間表現として扱う．本研究では，文レベルの中間表現とのアテンションを活用する方法として 2 つ提案する．

3.1 提案手法 1

提案手法 1 の概要図を図 2 に示す．提案手法 1 では，MirrorGAN[3] に倣い，文レベルの中間表現とのアテンション S_{attn} を次式 (7) のように計算する．

$$S_{attn} = \bar{h}_n \circ (softmax(h_j \circ \bar{h}_n)) \tag{7}$$

ここで，$\circ$ は要素積である．

そして，出力単語の確率分布は，S_{attn} を式 (4) に追加連結して算出した $\dot{h}$ に基づき獲得する．具体的には

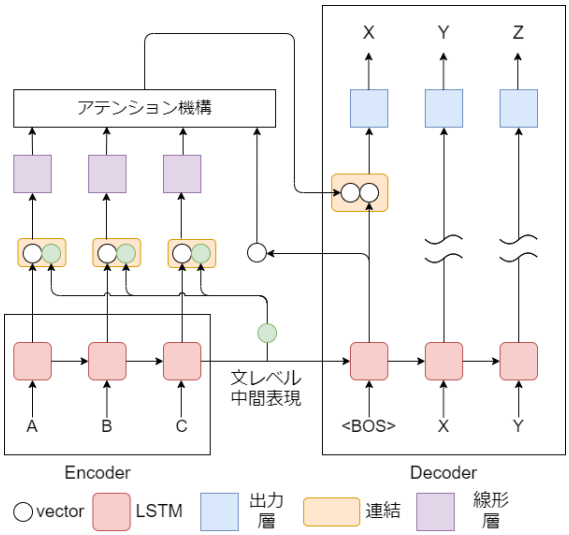

図 3: 提案モデル (提案手法 2)

次式 (8) のように算出した $\dot{h}$ を，式 (3) において h_j の代りに用いることで出力単語の確率分布を算出する．

$$\dot{h} = tanh(W_s[c_t; S_{attn}; h_j]) \tag{8}$$

ここで，$W_s \in R^{H \times 3H}$ は重み行列である．

3.2 提案手法 2

提案手法 2 の概要図を図 3 に示す．提案手法 2 のアテンション機構では，単語レベルの各中間表現に文レベルの中間表現を連結させたものとのアテンションを計算する．具体的には，エンコーダーの各隠れ状態とのアテンションを計算する際，次式 (9) のように算出した $\tilde{h}_i$ を使用する．つまり，式 (5) と式 (6) の $\bar{h}_i$ と $\bar{h}_k$ を $\tilde{h}_i$ や $\tilde{h}_k$ に置き換えて c_t を算出する．c_t の算出以降は従来のアテンション付き NMT と同様である．

$$\tilde{h}_i = W_l[\bar{h}_i; h_n] \tag{9}$$

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

4 実験

4.1 実験データ

本実験では，科学技術論文をもとに作成された日英対訳コーパスである Asian Scientific Paper Excerpt Corpus (ASPEC)[1]を用いた．英語文の単語分割には MOSES[2]を使用し，日本語文の単語分割には kytea[3]を使用した．学習データは日本語文と英語文ともに単語数が 50 単語以下である文対のみに制限し，10 万文対の対訳データを用いて学習を行った．開発データは 1,790 文対，テストデータは 1,812 文対である．

表 1: 実験結果

手法	BLEU(%)
ベースライン	21.26
提案手法 1	21.11
提案手法 2	21.98

4.2 実験設定

本実験では，4 節で提案した 2 つの提案手法と，ベースラインとして，文レベルの中間表現とのアテンションを考慮しない従来のアテンション機構付き NMT モデル [2] の翻訳性能を評価し，性能を比較した．

各モデルにおいて，エンコーダーには 2 層の双方向 LSTM を用い，デコーダーには 2 層の単方向 LSTM を用いた．単語埋め込み層，隠れ層の次元はともに 256 次元とした．各モデルの学習では，最適化手法として Adam[4] を使用し，learning rate=0.1 とした．また，ミニバッチサイズは 100 とし，p=0.1 の dropout を適用した．エポック数は 20 としてモデルを学習し，開発データにおいて最も BLEU[5] が高かったモデルを選択し，選択したモデルのテストデータに対する翻訳性能を最終的な評価とした．

5 結果

実験結果を表 1 に示す．表 1 より，提案手法 1 の翻訳性能はベースラインの翻訳性能を上回ることができなかったが，提案手法 2 の翻訳性能はベースラインの翻訳性能を上回った．この結果から，提案手法 2 のように文レベルの中間表現をアテンションに用いる手法は有効であることが実験的に確認できた．

[1]http://orchid.kuee.kyoto-u.ac.jp/ASPEC/

[2]https://github.com/moses-smt/mosesdecoder/tree/master/scripts/tokenizer

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

6 おわりに

本研究では，文レベルの中間表現とのアテンションをアテンション機構で用いるアテンション機構付き NMT モデルを提案した．そして，日英翻訳において，単語レベルの各中間表現に文レベルの中間表現を連結させたものとのアテンションを考慮することで翻訳性能を改善できることを確認した．今後は，ASPEC 日英翻訳以外の翻訳タスクにおいても提案モデルの有効性を検証したい．また，Transformer に基づく NMT モデルに対して，文レベルの中間表現とのアテンションを考慮するアテンション機構を導入したい．

7 謝辞

本研究成果は，国立研究開発法人情報通信研究機構の委託研究により得られたものである．また，本研究の一部は JSPS 科研費 18K18110 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] 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.

[2] Minh-Thang Luong, Hieu Pham, and Christopher D.Manning. Effective Approaches to Attention-based Neural Machine Translation. In Proceedings of Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 1412-1421, 2015.

[3] Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. MirrorGAN:Learning Text-to-image Generation by Redescription. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.1505-1514, 2019.

[4] Diederik Kingsma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In Proceedings of the 3rd International Conference for Learning Representations, 2015.

[5] 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, pp. 311-118, 2002.