マルチホップ注意機構を用いたニューラル機械翻訳*

飯田 頌平† 木村 龍一郎† 崔 鴻翌† 洪 博軒† 宇津呂 武仁† 永田 昌明‡
†筑波大学大学院 システム情報工学研究科 ‡NTT コミュニケーション科学基礎研究所

1 はじめに

ニューラル機械翻訳 (NMT) においては，モデル提案当初は Reccurent Neural Network (RNN) によるエンコーダ・デコーダモデル [1, 10] が主流であったが，近年は Transformer [15] や Convolutional Sequence-to-Sequence [6] といった非再帰型のエンコーダ・デコーダモデルの方がより優れた成果を挙げている．しかし RNN がすべての点で Transformer に劣るわけではない．例えば，Transformer は，パラメータの数を増やすことにより RNN の性能を改善したが，それにも関わらず，RNN と比較して長距離の依存関係に弱いことが知られている [13].

そこで本論文では，RNN に対して，Transformer とは異なる方式により高次のパラメータを持たせることによって，長距離の依存関係において Transformer よりも優れる手法を提案する．提案手法においては，マルチヘッド注意 [15] において，マルチホップ注意機構 [12] の要領で注意の繰り返し計算を行う．この機構を導入することにより，一つの入力に対するパラメータ数を増やすことができ，結果として，短文においても長文においても，ベースライン RNN の翻訳性能を改善するモデルを実現できた．そして，モデルの評価においては，通常の翻訳タスクによる評価に加えて，長文を対象として Transformer・ベースライン RNN との性能比較を行った．さらに，文脈情報を用いたニューラル機械翻訳の研究で用いられる 2-to-2 [14, 2] 手法を導入した文脈翻訳での性能評価を行った．ここでは，原言語文に対して，文脈情報として直前の原言語文一文を連結しており，通常よりも長い文長での翻訳性能評価となっている．

結果として，いずれの翻訳性能評価においても，提案手法により，ベースライン RNN による翻訳性能を有意水準 1%で改善することができた．また，ASPEC の日英・英日翻訳においては，単語数が 120〜129 となる長文において，Transformer による翻訳性能を有意水準 1%で改善することができた．これらの結果より，注意の計算を繰り返す提案手法の有用性が示された．

以下，2 節においては，RNN に基づく sequence-to-sequence における注意機構について述べ，本論文のマルチホップ注意機構を提案する．3 節においては，日英・英日機械翻訳タスクによる評価，および，文長毎の BLEU 評価を行い，長文での翻訳性能を示す．

2 RNN に基づく seq-to-seq モデルにおける注意機構

本節では，RNN に基づく sequence-to-sequence モデルにおける注意機構 (図 1(a)) について述べ，本論文のマルチホップ注意機構を提案する．そこでまず，ベースライン RNN [10]，マルチヘッド機構，階層型注意機構 [2]，および提案手法における入力文数，ヘッド，ホップ，および，ヘッド間相互作用についての比較結果を表 1 に示す．

2.1 マルチヘッド注意機構

本論文では，ヘッド数が N 本のマルチヘッド注意機構を実装する方法を次式で定義する．

$$s_i^{(k)} = W_a^{(k)} d_i \quad (1)$$

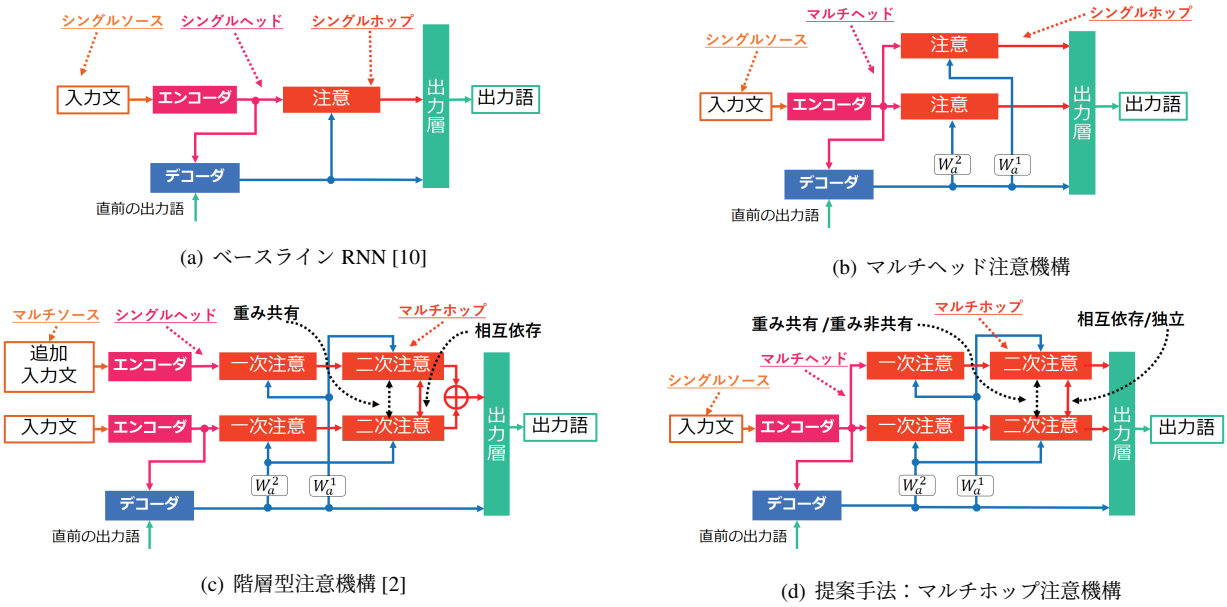
$$c_i^{(k)} = \mathrm{softmax}(s_i^{(k)} H^{\mathrm{T}}) H \quad (2)$$

式 (1) では，デコーダ RNN の内部状態 d_i をマルチヘッドに分割する．$W_a^{(k)}$ は学習可能なパラメータであり，これによって d_i が別々の値を指す $s_i^{(k)}$ に分割される．式 (2) では，各ヘッドに対して内積注意 [10, 15] によって，エンコーダ H と $k(=1,\ldots,N)$ 番目のヘッドのデコーダ $s_i^{(k)}$ との間の注意を求める．

[3] では Transformer の持つ様々な仕組みを RNN に取り入れる試みが行われ，その一つとして，RNN のソースターゲット注意機構に対してマルチヘッド注意機構が組み込まれた．本論文の提案手法との比較のために，[3] のマルチヘッド機構を単純化した模式図を図 1(b) に示す．

---

*A Study on Neural Machine Translation with Multi-Hop Attention
†Shohei Iida, Ryuichiro Kimura, Hongyi Cui, Po-Hsuan Hung, Takehito Utsuro, Graduate School of Systems and Information Engineering, University of Tsukuba
‡Masaaki Nagata, NTT Communication Science Laboratories, NTT Corporation, Japan



(a) ベースライン RNN [10]

(b) マルチヘッド注意機構

(c) 階層型注意機構 [2]

(d) 提案手法：マルチホップ注意機構

図 1: 比較対象モデル一覧

表 1: RNN の注意機構における入力文・ヘッド・ホップ・ヘッド間相互作用

手法	入力文	ヘッド	ホップ	ヘッド間相互作用
ベースライン RNN [10]	一文	シングル	シングル	なし
マルチヘッド注意機構	一文	マルチ	シングル	なし
階層型注意機構 [2]	二文	シングル	マルチ	重み共有; 相互依存
提案手法：マルチホップ注意機構	一文	マルチ	マルチ	重み共有/非共有; 相互依存/独立

2.2 相互依存型マルチホップ注意機構

end-to-end memory network [12] では注意を繰り返し計算するマルチホップ注意機構が用いられた．本論文では，原言語の入力文一文と原言語文の直前の一文を組み合わせる階層型注意機構 [2](図 1(c))[1] を参考にして，end-to-end memory network [12] のマルチホップ注意機構を RNN ベースの NMT モデルに導入し，相互依存型マルチホップ注意機構と呼ぶ．

$$e_i^{(k)} = v_b^{\mathrm{T}} \tanh(W_b s_i^{(k)} + U_b^{(k)} c_i^{(k)}) \quad (3)$$

$$\beta_i^{(k)} = \frac{\exp(e_i^{(k)})}{\sum_{n=1}^{N} \exp(e_i^{(n)})} \quad (4)$$

$$c_i'^{(k)} = \beta_i^{(k)} U_c^{(k)} c_i^{(k)} \quad (5)$$

具体的には，式 (3)，式 (4)，および，式 (5) によって各ヘッドのコンテキストベクトルの計算を行う (図 1(d))．

階層型注意機構では，複数の入力から得たエンコーダの出力値とデコーダの内部状態の間で注意を計算するため，各入力に対しては単一のヘッドが対応する．一方，本論文の提案手法では，単一の入力から得たエンコーダの出力値に複数のヘッドを当てることで入力の別々の箇所に注意が向くマルチヘッド注意機構をもとにして，各ヘッドがお互いに影響を及ぼし合いながら学習する．

式 (3) では，多層パーセプトロンを用いてデコーダ $s_i^{(k)}$ とヘッド $c_i^{(k)}$ の間の注意スコアを計算する (加法注意 [10])．ここで一次の注意の計算 (式 (2)) で用いた内積注意ではなく，加法注意を採用した理由は，各ヘッドのパラメータを共有することができるからである．式 (3) のパラメータ W_b と v_b はすべてのヘッドで共有されるため，各ヘッドが互いに影響を及ぼし合う作用を持つ [2]．内積注意は加法注意よりも優れるが [15]，複数のヘッド間で共有可能なパラメータを持つ設計ではなく，二次の注意の計算には不向きであるために不採用とした．式 (4) では，softmax 関数によって各ヘッドの注意スコアを $\beta_i^{(k)}$ に正規化し，式 (5) において，学習可能なパラメータ $U_c^{(k)}$ と $\beta_i^{(k)}$ によって，$c_i^{(k)}$ を新たなコンテキストベクトル $c_i'^{(k)}$ に更新する．最後に，得られた N 本のコンテキストベクトル $c_i'^{(k)}$ をデ

[1] 初出は [8] のマルチモーダル翻訳モデルである．

[2] 本節の「相互依存型」および次節の「独立型」とも，これらのパラメータの重みを二つのヘッド間で共有しない「重み非共有」設定との間で性能比較した結果においては，総合的には，翻訳精度において「重み共有」が「重み非共有」を上回る傾向にあった．

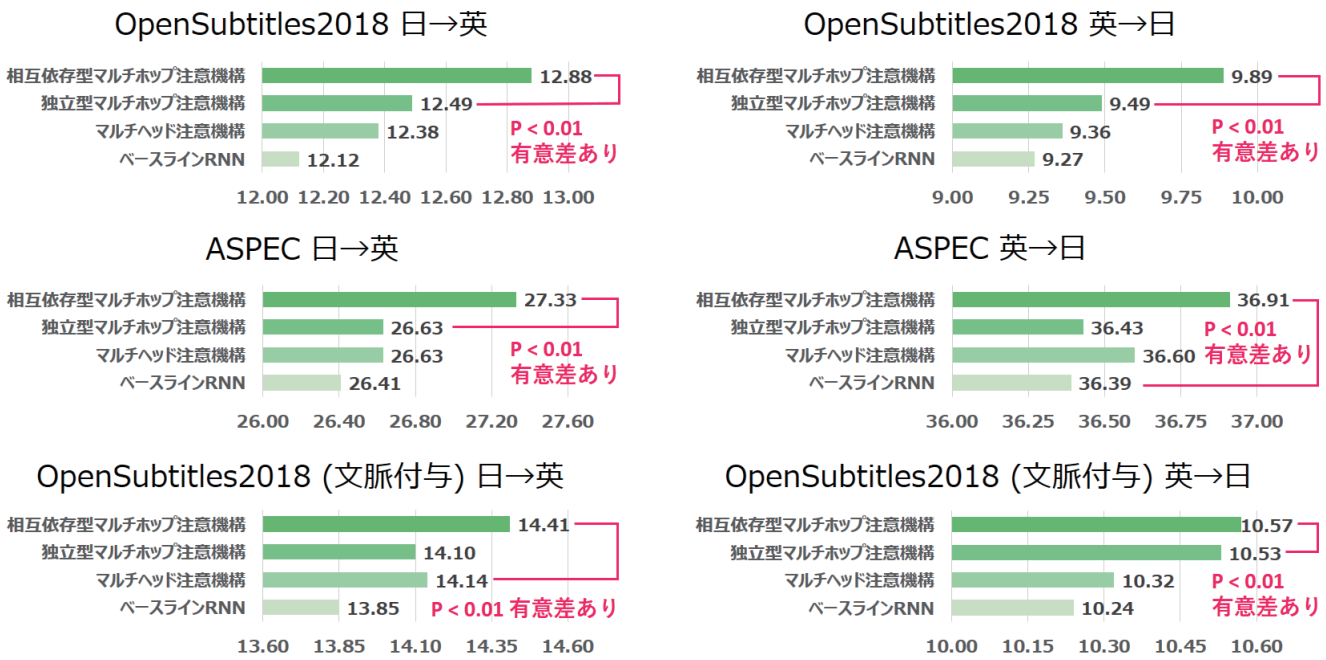


図 2: BLEU 評価結果

コーダの内部状態 d_i と連結させて，単語の予測分布 $p(y_i|y_{i-1}, X)$ を得る．

$$o_i = \tanh(W_o[d_i; c_i'^{(1)}; ...; c_i'^{(k)}]) \quad (6)$$

$$p(y_i|y_{i-1}, X) = \mathrm{softmax}(o_i) \quad (7)$$

ここで，W_o は学習可能なパラメータである．

2.3 独立型マルチホップ注意機構

相互依存型マルチホップ注意機構では，二次の注意 $c_i'^{(k)}$ を計算する際 (式 (5)) に，他のヘッドの情報を使い，また加法注意のパラメータ W_b, v_b をすべてのヘッドで共有する (式 (3))．一方，本節で述べる独立型マルチホップ注意機構では，加法注意を使わずに多層パーセプトロンのパラメータ $U_c^{(k)}$ のみで二次の注意を計算する．この方式では，式 (5) を次式に変更する．

$$c_i'^{(k)} = U_c^{(k)} c_i^{(k)} \quad (8)$$

この場合，加法注意がないため複数のヘッド間で共有するパラメータがない．また，softmax 関数を用いて各ヘッドを正規化することをしないため，二次の注意の計算において同一ヘッドの情報しか用いない．

3 評価

評価尺度として BLEU を用いて，日英・英日機械翻訳タスクによって提案手法の有用性を評価した．

3.1 データセット

OpenSubtitles2018 [9] および科学技術論文コーパス Asian Scientific Paper Excerpt Corpus (ASPEC) [11] から取得した日英対訳コーパスを利用した．ASPEC においては，訓練文 3,000,000 文のうち，文アライメントが上位の 1,000,000 文を使用した．

3.2 実験設定

ベースラインとして RNN に基づく sequence-to-sequence モデル [10] を用いる．埋め込み層を 512 次元，隠れ層を 1,024 次元，エンコーダ・デコーダをそれぞれ 1 層ずつとする．Transformer においては，ヘッド数は fairseq [7] のデフォルト設定 [3] に従い 4 本とした．すべてのモデルで 20epoch の訓練を行った．

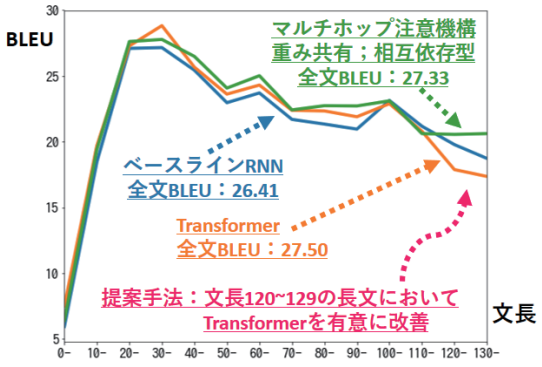
3.3 評価結果

評価結果を図 2 に示す．提案手法では相互依存型 (重み共有) マルチホップ注意機構が最も高い性能を示し，いずれのデータセットにおいてもベースライン RNN，マルチヘッド注意機構，独立型 (重み共有) マルチホップ注意機構の BLEU スコアを上回り，ASPEC の英日翻訳以外では他のいずれの手法とも有意水準 1%での有意差を達成した．このことから，相互依存型マルチホップ注意機構によってエンコーダとデコーダを接続する方式が，独立型よりも有効であることが判明した [4]．

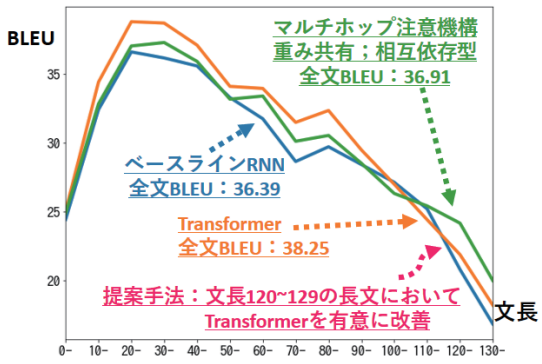
一方，ベースライン RNN，Transformer との間で ASPEC における文長ごとの BLEU の比較をした評価結果

[3] 埋め込み層・隠れ層を各 512 次元，エンコーダ・デコーダを各 6 層とした．

[4] 相互依存型において複数ヘッド間でパラメータを共有しない場合でも，本節における相互依存型とほぼ同等の性能となった．



(a) 日英方向 (提案手法と Transformer の間で有意差なし)



(b) 英日方向 (提案手法と Transformer の間で有意差あり)

図 3: 文長毎の BLEU (ASPEC)

を図 3 に示す．全文長での BLEU では Transformer に劣るものの，日英翻訳においては有意差はなく，RNN ベースでありながら Transformer に匹敵する性能のモデルとなった．また，単語数が 120〜129 となる長文については，有意水準 1%で Transformer の BLEU を有意に改善した．

4 関連研究

長距離の依存関係に弱いという Transformer の弱点を克服する取り組みとしては，Universal Transformer [4] が挙げられる．Universal Transformer の仕組みにおいては，Transformer の各層のパラメータを共有させた状態で注意計算を繰り返す一方で，Transformer よりもパラメータ数が増加する点が弱点である．その他，Transformer のパラメータ数を増やすことによりモデルを性能を改善する試みとして BERT [5] が知られている．一方，提案手法においては，Transformer よりもパラメータ数が少ないためにモデルが軽い点が長所となる．

また，翻訳以外のタスク，例えば質問応答では，回答の情報源となるパッセージと質問文から構成される複数入力の間で異なる注意を計算し，応答文を生成するモデルが知られている．その中でも [16] では，複数入力間での注意を計算するマルチホップ注意機構を用いており，この点において本論文と関連している．

5 おわりに

本論文では，RNN ベース NMT モデルにおいて，マルチヘッド注意機構の各ヘッドが相互に依存するマルチホップ注意機構を提案した．提案手法を RNN のソースターゲット注意機構に適用することにより，特に長文における翻訳精度が改善し，Transformer の翻訳精度を有意に上回った．今後の課題として，提案手法を Transformer のマルチヘッド注意機構へ適用し，両者の長所を組み合わせた手法の開発が挙げられる．

参考文献

[1] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In *Proc. 3rd ICLR*, 2015.

[2] R. Bawden, R. Sennrich, A. Birch, and B. Haddow. Evaluating discourse phenomena in neural machine translation. In *Proc. NAACL-HLT*, pp. 1304–1313, 2018.

[3] M. X. Chen, et al. The best of both worlds: Combining recent advances in neural machine translation. In *Proc. 56th ACL*, pp. 76–86, 2018.

[4] M. Dehghani, S. Gouws, O. Vinyals, J. Uszkoreit, and Ł. Kaiser. Universal transformers. *arXiv preprint arXiv:1807.03819*, 2018.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *CoRR*, Vol. abs/1810.04805, 2018.

[6] J. Gehring, M. Auli, D. Grangier, and Y. Dauphin. A convolutional encoder model for neural machine translation. In *Proc. 55th ACL*, pp. 123–135, 2017.

[7] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin. Convolutional sequence to sequence learning. In *Proc. 34th ICML*, pp. 1243–1252, 2017.

[8] J. Libovický and J. Helcl. Attention strategies for multi-source sequence-to-sequence learning. In *Proc. 55th ACL*, pp. 196–202, 2017.

[9] P. Lison, J. Tiedemann, and M. Kouylekov. Opensubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proc. 11th LREC*, pp. 1742–1748, May 7-12, 2018 2018.

[10] M. Luong, H. Pham, and C. D. Manning. Effective approaches to attention-based neural machine translation. In *Proc. EMNLP*, pp. 1412–1421, 2015.

[11] T. Nakazawa, M. Yaguchi, K. Uchimoto, M. Utiyama, E. Sumita, S. Kurohashi, and H. Isahara. ASPEC: Asian scientific paper excerpt corpus. In *Proc. 10th LREC*, pp. 2204–2208, 2016.

[12] S. Sukhbaatar, A. Szlam, J. Weston, and R. Fergus. End-to-end memory networks. In *Proc. 28th NIPS*, pp. 2440–2448. 2015.

[13] G. Tang, M. Müller, A. Rios, and R. Sennrich. Why self-attention? a targeted evaluation of neural machine translation architectures. In *Proc. EMNLP*, pp. 4263–4272, 2018.

[14] J. Tiedemann and Y. Scherrer. Neural machine translation with extended context. In *Proc. 3rd DiscoMT*, pp. 82–92, 2017.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *Proc. 30th NIPS*, pp. 5998–6008, 2017.

[16] C. Xiong, V. Zhong, and R. Socher. DCN+: mixed objective and deep residual coattention for question answering. *Proc. 6th, ICLR*, 2018.