係り受け構造に対する相対位置表現を考慮したNMT

表　悠太郎[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
{omote@ai., tamura@, ninomiya@}cs.ehime-u.ac.jp

1　はじめに

機械翻訳は自然言語処理の初期から盛んに研究され，様々な手法が提案されてきているが，近年では，ニューラルネットワークを用いた機械翻訳（NMT）が高い精度を実現しており，主流となっている．NMT の中でも，特に，Self Attention を用いた Transformer[1] が state-of-the-art の精度を達成し，注目を集めている．

これまで，統計的機械翻訳や Recurrent Neural Network（RNN）に基づく NMT 等では，原言語文や目的言語文，あるいはその両方の構文情報（句構造や係り受け構造など）を活用することで翻訳精度の改善が行われてきた．しかしながら，Transformer の翻訳モデルでは，これまで構文情報は陽に活用されていない．そこで本研究では，Transformer で構文情報を活用することにより，その翻訳精度の改善を試みる．特に，本研究では，Transformer における構文情報の活用の第一歩として，原言語文の係り受け構造の活用に取り組む．

初期の Transformer[1] では，単語の絶対的な位置を Positional Encoding を用いてエンコードすることで，各単語の文中における位置情報を考慮する．Shaw ら [2] は，この単語の絶対的な位置情報に加えて，単語間の文中における相対的な位置関係を Self Attention で考慮することで Transformer の精度改善を行っている．本研究では，Shaw ら [2] に倣い，原言語文を係り受け解析した結果得られる係り受け木における単語間の相対的位置関係を埋め込んだベクトルを Transformer エンコーダ内の Self Attention に付加することで，原言語側の係り受け構造を Transformer で活用する手法を提案する．ASPEC の英日翻訳タスク [3] において，原言語文の係り受け構造を相対位置表現で考慮する提案モデルは従来の Transformer モデル [1, 2] と同等かそれ以上の翻訳精度を達成できることを示す．

2　従来手法

2.1　Transformer

Transformer は，エンコーダレイヤとデコーダレイヤがそれぞれ複数層スタックされたエンコーダ･デコーダ構造を持つ．エンコーダレイヤは，入力に近い方から順に，Self Attention，位置ごとのフィードフォーワードネットワーク（FFN）の 2 つのサブレイヤから構成されている．デコーダレイヤは，下位のサブレイヤから順にマスキング付き Self Attention，原言語文と目的言語文間の Attention（Src-Target Attention），位置ごとの FFN の 3 つのサブレイヤーから構成されている．各サブレイヤ間では，残差接続を行った後に Layer Normalization が適用される．つまり，下位のサブレイヤからの出力を $\boldsymbol{x}$ としたとき，$LayerNorm(\boldsymbol{x} + SubLayer(\boldsymbol{x}))$ がサブレイヤの出力となる．

Self Attention と Src-Target Attention は Multi-Head Attention を用いて実現されている．Multi-Head Attention では，まず，3 つの入力ベクトル $\boldsymbol{q}, \boldsymbol{k}, \boldsymbol{v} \in \mathbb{R}^{d_{model}}$ を重み行列 $W_i^Q, W_i^K, W_i^V \in \mathbb{R}^{d_{model} \times d_k} (i = 1, \ldots, h)$ により，d_{model} 次元から d_k 次元に線形写像した後，h 個の内積 Attention を計算する．ここで，d_{model} は元々の入力ベクトルの埋め込み次元であり，$d_k = d_{model}/h$ である．また，それぞれの内積 Attention をヘッド（$Head_i$ $(i = 1, \ldots, h)$）と呼ぶ．

$$Head_i = Attention(\boldsymbol{q}W_i^Q, \boldsymbol{k}W_i^K, \boldsymbol{v}W_i^V) \quad (1)$$

$$Attention(\boldsymbol{q}, \boldsymbol{k}, \boldsymbol{v}) = softmax(\frac{\boldsymbol{q}\boldsymbol{k}^T}{\sqrt{d_k}})\boldsymbol{v} \quad (2)$$

その後，各ヘッドを連結した後，重み行列 $W^o \in \mathbb{R}^{d_{model} \times d_{model}}$ で線形写像する機構が Multi-Head Attention である．

$$MultiHead(\boldsymbol{q}, \boldsymbol{k}, \boldsymbol{v}) = Concat(Head_1, \ldots, Head_h)W^O \quad (3)$$

FFN は入力 $\boldsymbol{x}$ に対して，以下の計算を行う．

$$FFN(\boldsymbol{x}) = max(0, \boldsymbol{x}W_1 + \boldsymbol{b}_1)W_2 + \boldsymbol{b}_2 \quad (4)$$

ここで，$W_1 \in \mathbb{R}^{d_{model} \times d_{ff}}, W_2 \in \mathbb{R}^{d_{ff} \times d_{model}}$ は重み行列，$\boldsymbol{b}_1, \boldsymbol{b}_2$ はバイアス項である．

Transformer は，RNN に基づく NMT とは異なり再帰的な構造を持たないため，単語の系列情報を付与する必要がある．そこで，Transformer では，入力文中の各単語の埋め込み表現行列に対して，文中の絶対的な単語位置情報をエンコードした行列 PE を加えたものをエンコーダやデコーダの入力とする． PE の各成分は異なる周波数の sin,cos 関数を用いて次式により算出される．

$$PE(pos, 2i) = \sin(pos/10000^{2i/d_{model}}) \quad (5)$$

$$PE(pos, 2i+1) = \cos(pos/10000^{2i/d_{model}}) \quad (6)$$

ここで，pos は単語の位置，i は各成分の次元を表す．

2.2 Self Attention

Self Attention は，式 (3) の $\boldsymbol{q}, \boldsymbol{k}, \boldsymbol{v}$ のすべてに単語の埋め込みベクトル系列 $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ を代入し計算を行う．具体的には，各ヘッドでは以下のような荷重和を計算する．

$$\boldsymbol{z}_i = \sum_{j=1}^{n} \alpha_{ij} \boldsymbol{x}_j W^V \quad (7)$$

ここで，$\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ が Self Attention の出力系列である．重み係数 α_{ij} はソフトマックス関数を用いて以下の通り計算される．

$$\alpha_{ij} = \frac{\exp(\boldsymbol{e}_{ij})}{\sum_{k=1}^{n} \exp(\boldsymbol{e}_{ik})} \quad (8)$$

また，$\boldsymbol{e}_{ij}$ は以下のようにして計算される．

$$\boldsymbol{e}_{ij} = \frac{(\boldsymbol{x}_i W^Q)(\boldsymbol{x}_j W^K)^T}{\sqrt{d_z}} \quad (9)$$

2.3 相対的位置を考慮した Self Attention

Shaw ら [2] は 2 単語間の文における相対的な位置関係を Self Attention で捉える手法を提案した．Shaw ら [2] の手法では，入力文中の各単語の中間表現 $\boldsymbol{x}_i$,$\boldsymbol{x}_j$ 間の関係はベクトル $\boldsymbol{a}_{ij}^V, \boldsymbol{a}_{ij}^K \in \mathbb{R}^{d_k}$ で表現する．そして，サブレイヤの出力に単語間の相対位置情報を付加して次の層への入力とする．具体的には，式 (7) の代わりに次式を用いる．

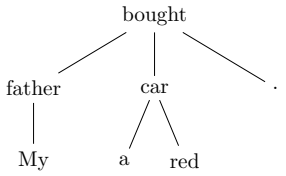

図 1: 係り受け木の例

$$\boldsymbol{z}_i = \sum_{j=1}^{n} \alpha_{ij} (\boldsymbol{x}_j W^V + \boldsymbol{a}_{ij}^V) \quad (10)$$

また，self attention 計算過程の $\boldsymbol{e}_{ij}$ 算出時にも単語間の相対位置情報を考慮するため，式 (9) の代わりに次式を用いる．

$$\boldsymbol{e}_{ij} = \frac{\boldsymbol{x}_i W^Q (\boldsymbol{x}_j W^K + \boldsymbol{a}_{ij}^K)^T}{\sqrt{d_z}} \quad (11)$$

ここで Shaw ら [2] は，単語間が一定距離以上離れると離れ具合の影響は少ないと仮定し，相対的位置の最大値を定数 k と定め，それより離れた相対位置は最大値 k とした．また，文中のある単語から後ろを正の方向，前を負の方向と考え，二単語間の相対的位置関係は，以下の通り，$2k+1$ 個のユニークなラベルで捉える．

$$\boldsymbol{a}_{ij}^K = \boldsymbol{w}_{clip(j-i,k)}^K \quad (12)$$

$$\boldsymbol{a}_{ij}^V = \boldsymbol{w}_{clip(j-i,k)}^V \quad (13)$$

$$clip(x, k) = max(-k, min(k, x)) \quad (14)$$

ここで，各相対位置ラベルの埋め込み表現は，$\boldsymbol{w}^K = (\boldsymbol{w}_{-k}^K, \ldots, \boldsymbol{w}_k^K)$ と $\boldsymbol{w}^V = (\boldsymbol{w}_{-k}^V, \ldots, \boldsymbol{w}_k^V)$ $(\boldsymbol{w}_k^V, \boldsymbol{w}_k^K \in \mathbb{R}^{d_k})$ であり，学習されるパラメータである．

3 提案手法

3.1 係り受け構造に対する相対位置

提案手法では，原言語文の係り受け構造を Transformer で利用する．係り受けとは，単語間の「修飾」「被修飾」の関係のことであり，方向性を持つ．“My father bought a red car.” という文の係り受け構造を表す係り受け木を図 1 に示す．係り受け木では，単語 A が単語 B を修飾する関係を，単語 A を単語 B の子ノードにすることで表現する．提案手法では，文中の 2 単語 x_i, x_j に対して，係り受け木における相対的位置関係を表す位置ラベル $label_{i,j}$ を，次の通り決定する．

表 1: 係り受け構造に対する相対位置ラベルの例

	My	father	bought	a	red	car	.
My	$self$	-1	-2	non_dep	non_dep	non_dep	non_dep
father	1	$self$	-1	non_dep	non_dep	sib	sib
bought	2	1	$self$	2	2	1	1
a	non_dep	non_dep	-2	$self$	sib	-1	non_dep
red	non_dep	non_dep	-2	sib	$self$	-1	non_dep
car	non_dep	sib	-1	1	1	$self$	sib
.	non_dep	sib	-1	non_dep	non_dep	sib	$self$

- 単語 x_i と単語 x_j に対応するノード n_i とノード n_j が祖先子孫関係の場合，$label_{ij} = depth(n_j) - depth(n_i)$ とする．ここで，$depth(n)$ はノード n の深さを表す．例えば，図 1 において $label_{My,bought} = 2 - 0 = 2$ である．この定義より，係り受け木においてある単語の親方向は正，子方向は負の値で相対的位置関係を表すことができる．
- 単語 x_i と単語 x_j に対応するノード n_i とノード n_j が兄弟関係にある場合，兄弟関係ラベル sib（$label_{ij} = sib$）とする．ここで，2 ノードが兄弟関係の場合，祖先子孫関係とは異なり，方向性や距離の情報は相対的位置ラベルに持たせないことに注意されたい．
- 単語自身との相対的位置ラベルは自分自身を表すラベル $self$ とする．つまり，任意の単語 x_i に対して，$label_{ii} = self$ である．
- 上記 3 パタン以外の 2 単語間の相対的位置ラベルは，依存関係なしを表すラベル non_dep とする．

図 1 の係り受け木における 2 単語間の相対的位置ラベルを表 1 に示す．

3.2 係り受け構造に対する相対位置表現を用いた Self Attention

提案手法では，Shaw ら [2] に倣い，3.1 節で特定した係り受け木における相対的位置ラベルを埋め込みベクトルで表現し，2 単語間の係り受け木における相対的位置の情報を Self Attention で考慮する．具体的には，原言語文の各単語の中間表現 $\boldsymbol{x}_i$,$\boldsymbol{x}_j$ 間の係り受け構造に対する相対位置関係をベクトル $\boldsymbol{b}^V_{ij}, \boldsymbol{b}^K_{ij} \in \mathbb{R}^{d_k}$ で表現し，式 (10)，(11) の代わりに以下を用いる．

$$\boldsymbol{z}_i = \sum_{j=1}^{n} \alpha_{ij}(\boldsymbol{x}_j W^V + \boldsymbol{b}^V_{ij}) \tag{15}$$

$$\boldsymbol{e}_{ij} = \frac{\boldsymbol{x}_i W^Q (\boldsymbol{x}_j W^K + \boldsymbol{b}^K_{ij})^T}{\sqrt{d_z}} \tag{16}$$

係り受け木における相対位置関係においては，一定以上距離が大きいと単語間の依存関係は少なくなると仮定し，最大距離 k 以上の離れた単語間の相対位置は Self Attention で考慮しないようにする．先祖・子孫関係に 2.3 節と同様に定数 k で最大距離に制限をかけ，最大距離以上の位置の単語は Self Attention で考慮しないようにする．したがって，原言語文中の各単語の中間表現 $\boldsymbol{x}_i$，$\boldsymbol{x}_j$ に対して，係り受け木における相対的位置関係を埋め込んだベクトル $\boldsymbol{b}^V_{ij}, \boldsymbol{b}^K_{ij}$ は次式で表現できる．

$$\boldsymbol{b}^{\boldsymbol{K}}_{\boldsymbol{ij}} = \begin{cases} \boldsymbol{w}^{\boldsymbol{K}}_{\boldsymbol{label_{ij}}} & (label_{ij} = sib, self, \\ & |label_{ij}| <= k \text{ の時}) \\ \boldsymbol{0} & (otherwise) \end{cases}$$

$$\boldsymbol{b}^{\boldsymbol{K}}_{\boldsymbol{ij}} = \begin{cases} \boldsymbol{w}^{\boldsymbol{V}}_{\boldsymbol{label_{ij}}} & (label_{ij} = sib, self, \\ & |label_{ij}| <= k \text{ の時}) \\ \boldsymbol{0} & (otherwise) \end{cases}$$

以上のように，式 (15)，(16) を用いて，Self Attention で文内における相対位置表現の代わりに係り受け構造に対する相対位置表現を考慮する手法を提案手法 1 とする．

また，提案手法 1 に加えて，係り受け構造に対する相対位置表現と 2.3 節の文内における相対位置表現の両方の情報を考慮する手法を提案手法 2 として提案する．具体的には，$\boldsymbol{a}^V_{ij}, \boldsymbol{b}^V_{ij}$ と $\boldsymbol{a}^K_{ij}, \boldsymbol{b}^K_{ij}$ をそれぞれ結合し，重み行列 $W^V_{rel}, W^K_{rel} \in \mathbb{R}^{2d_k \times d_k}$ を用いて線形変換を施したベクトル $\boldsymbol{c}^V_{ij}, \boldsymbol{c}^K_{ij} \in \mathbb{R}^{d_k}$ を 2 単語間の相対位置情報として用いる．つまり，提案手法 2 では，式 (15)，

表 2: 実験結果

モデル	BLEU
$Trans_{abs}$[1]	30.21
$Trans_{rel}$[2]	30.59
提案手法 1	29.77
提案手法 2	30.63

(16) の代わりに次式 (19)，(20) を用いる．

$$\boldsymbol{c}_{ij}^{V} = Concat(\boldsymbol{a}_{ij}^{V}, \boldsymbol{b}_{ij}^{V})W_{rel}^{V} \tag{17}$$

$$\boldsymbol{c}_{ij}^{K} = Concat(\boldsymbol{a}_{ij}^{K}, \boldsymbol{b}_{ij}^{K})W_{rel}^{K} \tag{18}$$

$$\boldsymbol{z}_i = \sum_{j=1}^{n} \alpha_{ij}(\boldsymbol{x}_j W^{V} + \boldsymbol{c}_{ij}^{V}) \tag{19}$$

$$\boldsymbol{e}_{ij} = \frac{\boldsymbol{x}_i W^{Q}(\boldsymbol{x}_j W^{K} + \boldsymbol{c}_{ij}^{K})^{T}}{\sqrt{d_z}} \tag{20}$$

4 実験

本節では，ASPEC[3] データセットを用いた英日翻訳タスクで提案手法の有効性を検証する．訓練データとして 100,000 文対，検証データとして 1,790 文対，テストデータとして 1,812 文対を用いた．英語文は Stanford CoreNLP[4] を用いてトークン化及び係り受け解析を行った．日本語文は KyTea[5] を用いてトークン化した．実験では，3 節で提案した 2 種類の提案手法を，従来の絶対的位置表現を考慮した Transformer（$Trans_{abs}$）[1] と相対的位置表現を考慮した Transformer（$Trans_{rel}$）[2] と比較する．評価する全手法の Transformer のハイパーパラメータは Vaswani ら [1] の設定に倣い，エンコーダ及びデコーダレイヤのスタック数は 6，ヘッド数は 8, 埋め込み次元は 512 次元とした．また optimizer は Adam を用い，$\beta_1 = 0.9, \beta_2 = 0.98, \epsilon = 10^{-9}$ と設定した．また，提案手法 1,2 及び $Trans_{rel}$ において，考慮する相対的位置の最大距離は $k = 2$ とした．ミニバッチサイズは 100，エポック数は 50 とし，検証データに対して最も精度が良かったエポックのモデルをテストデータに適用して翻訳性能を求めた． 翻訳性能は BLEU で評価した．

評価結果を表 2 に示す．表 2 より，係り受け木に対する相対的位置表現のみを用いる提案手法 1 は $Trans_{abs}$ よりも 0.44 ポイント BLEU が低くなったが， 文における相対的位置表現及び係り受け木に対する相対的位置表現の両方を用いる提案手法 2 は，$Trans_{abs}$ と比較して 0.42 ポイント BLEU が上回った．また，提案手法 2 は，$Trans_{rel}$ とは 0.04 ポイントの差となり，同等の性能であった．

5 おわりに

本研究では，Transformer において原言語文の係り受け構造を活用するため，原言語の係り受け木における単語間の相対的位置関係を Self Attention の中の相対的位置表現で考慮する手法を提案した．評価実験を通じて，提案モデルが従来の Transformer モデル [1, 2] と同等かそれ以上の翻訳精度を達成できることを確認した．今後は，係り受け木から相対的位置情報を抽出する手法を改善したり，他のデータセットや言語対での評価を行うなどしていきたい．

6 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 25280084 及び 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] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. Attention is all you need. In *Proc. of NIPS 2017*.

[2] P. Shaw, J. Uszkoreit, and A. Vaswani. Self-attention with relative position representations. In *Proc. of NAACL HLT 2018 (Short Papers)*.

[3] T. Nakazawa, M. Yaguchi, K. Uchimoto, M. Utiyama, E. Sumita, S. Kurohashi, and H. Isahara. Aspec: Asian scientific paper excerpt corpus. In *Proc. of LREC 2016*.

[4] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky. The Stanford CoreNLP natural language processing toolkit.

[5] G. Neubig, Y. Nakata, and S. Mori. Pointwise prediction for robust, adaptable japanese morphological analysis. In *Proc. of ACL HLT 2011*.