Vol.2017-NL-233 No.3
2017/10/24

サブワードユニットを用いたニューラル機械翻訳における形態素情報の効果

中村 尚道[1,a)]　井佐原 均[2,b)]

概要：深層学習は自然言語処理などの様々な分野において，それまでの研究を上回る成果を出している．機械翻訳分野においても，既存の統計的機械翻訳より高い性能を得られることが報告されている．しかしながら，ニューラル機械翻訳は大量のコーパスと高い計算コストを必要とする．計算コストを削減するために，既存研究では語彙内の低頻度な語句を記号やタグなどに置換する手法が主に用いられている．しかしながら，この手法は文の意味を曖昧にし，翻訳の性能を低下させることも報告されている．この問題を解決するためにバイト対符号化や Wordpiece Model などの手法を用いたサブワードユニットが提案されている．これらの手法は予め指定された語彙数から語彙を作成できるため，意味を曖昧にすることなく文を分割することができる．また，これらの手法は文を意味を持たないトークンに分解するため，入力列はトークンの集合となる．これはニューラル機械翻訳と相性が良く，翻訳精度を向上させることが報告されている．この結果から，ニューラル機械翻訳において言語学的な情報は必ずしも必要では無いとも考えられるが，我々はサブワードユニットに対して形態素情報を付与することで，翻訳精度が向上することを示した．サブワードユニットに対しても言語学的な情報が有用といえる．

Effect of Linguistic information in Nueral Machine Translation

Nakamura Naomichi[1,a)]　Isahara Hitoshi[2,b)]

1. はじめに

ニューラル機械翻訳は深層学習を用いた機械翻訳であり，近年急速に発展している．また，ニューラル機械翻訳はそれまでのフレーズベースの統計的機械翻訳に比べて高い性能を得られることが報告されている．再帰型ニューラルネットワークはニューラルネットワークのモデルの一つであり，連続的に可変長の入力を行うことができる．これは，入力の長さが一定でない自然言語処理や音声言語処理などのタスクに対して有用である．加えて，Long-Short Term Memory（LSTM）[6] により，それまでのニューラル機械翻訳において長文を正しく翻訳できないという課題を解決することができた．また，符号化復号化モデル（Encoder-Decoder model）[1] および系列変換モデル（Sequence-to-Sequence model）[14] は，翻訳や質問応答，キャプション生成などのエンドツーエンドであるタスクにおいて高い性能を発揮することが報告された．

しかしながら，ニューラル機械翻訳による翻訳には訳抜けや重複，低頻度語による誤訳などの問題が存在する．ニューラル機械翻訳は膨大な量のコーパスから学習しており，膨大な量のコーパスには単語が多く含まれているため語彙数が膨大になる．膨大な語彙数は単語分散表現の次元数を増加させるため，計算時間が増加する．ニューラル機械翻訳において学習時間の増加は深刻な問題であり，高い精度を記録した既存研究では 24 ギガバイトのメモリをもつ 96 枚の GPU で 6 日の学習時間を必要としたことが報告されている [15]．この問題を解決するために，近年のモデルでは低頻度な語彙を<unk>などの記号に置換する手法が用いられている．しかしながら，この手法は以下に示すように文を曖昧にすることが考えられる．

(1) Mike chases the pet with *mottle*.

[1] 豊橋技術科学大学 情報・知能工学専攻
[2] 豊橋技術科学大学 情報メディア基盤センター
[a)] n143361@edu.tut.ac.jp
[b)] isahara@tut.jp

(2) Mike chases the pet with *scooter*.
(3) Mike chases the pet with *Sullivan*.

これらの三つの文章において，末尾の単語を全て記号に置き換えた場合，全ての文が同一の意味を持つ文となる．この低頻度語問題を解決するために，サブワードユニットを用いた手法が報告されている [13]．この手法ではデータ圧縮法の一つであるバイト対符号化 [5] を用いて文のトークン化を行う．トークナイザとしてのバイト対符号化は文を文字に分解し，予め指定したサイズに到達するまで文字の結合を繰り返すことでトークン化を行う．したがって，未知語は発生しないが，単語は破壊され，意味を持たないトークンへと変換される．既存研究では、サブワードユニットを用いた Wordpiece モデルにより高い翻訳精度を得られることが報告されている [15]．意味を持たないサブワードユニットが翻訳精度を上昇させたことから，ニューラル機械翻訳において言語学的な情報は必ずしも必要ではないと考えられる．そこで，我々はサブワードユニットに対して形態素情報を付与することで，ニューラル機械翻訳における言語学的な情報の効果を評価する．

2. 関連研究

機械翻訳分野において，ニューラル機械翻訳はフレーズベース統計的機械翻訳より高い性能を得られることが多くの研究で報告されている．その性能は、符号化復号化モデル [1] の提案により，急速に上昇している．近年では，さらに性能を向上させるために様々なモデルが報告されている．トピックモデルをニューラル機械翻訳に導入することで，ドキュメントが持つ情報をモデルに付与するモデル [16] や，予めフレーズベース統計的機械翻訳によって翻訳した文を入力として用いるモデル [11] なども提案されている．

ニューラル機械翻訳には文を曖昧にする低頻度語問題が存在する．これを解決するために低頻度語を語彙内の高頻度な単語に置換する手法が報告されている [8]．しかし，置換された単語を元の単語へ再置換する際に失敗し，翻訳精度を低下させてしまうことも分かっている．また，入力として単語の代わりに文字を用いる手法 [2] も提案されているが，単語を文字に分割してしまうため入力長が長大になり，文脈が失われやすい．ニューラルネットの構造として構文木を用いることで文の構造をニューラル機械翻訳に導入する手法も報告されている [4]．これらの研究では入力として単語を用いているため未知語に対して頑健でない．そのため，本研究ではサブワードを用いることで未知語による誤訳を抑制し，かつ言語学的情報を導入することでサブワードを用いたモデルでの翻訳能力向上を試みる．

3. ニューラル機械翻訳

3.1 系列変換モデル

系列変換モデル [14] は，入力列を隠れ状態に変換する符号化器と，隠れ状態を出力列へ変換する復号化器から構成されている．機械翻訳分野においては，符号化器は翻訳元言語の文を隠れ状態へと変換し，復号化器は隠れ状態から翻訳先言語の文を生成する．このモデルでは，二つの言語を文の終端を表す特殊な記号（一般的には<eos>が用いられる）で接続する．出力列は翻訳先単語列 $S = (s_1, s_2, \cdots, s_m)$ および翻訳先単語列 $T = (t_1, t_2, \cdots, t_n)$ を用いて以下のように表される．

$$p(T|S) = \prod_{i=1}^{n} p(t_i|t_{<i}, S) \tag{1}$$

このとき，$p(t_i|t < i, S)$ はステップ i における単語 t_i を出力する条件付き確率である．またこの条件確率は，$softmax$ 関数を用いて以下のように表すこともできる．

$$p(t_i|t < i, S) = softmax(f(h_i)) \tag{2}$$

式 2 において，h_i は隠れ状態であり，f は隠れ状態を単語分散表現の次元数へと変換する関数である．ここで，隠れ状態 h_i は以下の式から算出される．

$$h_i = g(h_{i-1}, \mathbf{s}) \tag{3}$$

h_{i-1} は直前のステップにおける隠れ状態であり，$\mathbf{s}$ は符号化器から出力される入力列を表す隠れ状態である．関数 g は隠れ状態を算出する再帰型ニューラルネットワークユニットであり，LSTM や GRU などが用いられる．

3.2 注意機構

注意機構（Attention mechanism）はより長い入力列に対しても学習を可能にするための手法である [3][9]．また，入力列と出力列のアライメントを学習することもできる．ニューラル機械翻訳は文レベルで翻訳を行うため，単語レベルで翻訳を行うフレーズベース統計的機械翻訳に比べて翻訳元文と翻訳先文のアライメントが不明確であった．注意機構は，以下の手法を用いてこの問題を解決することができる．また，本研究では [9] において提案されている手法を実装している．注意機構を導入した隠れ状態 $\tilde{h_i}$ を式 2 に導入すれば，条件付き確率は以下のように表される．

$$p(t_i|t < i, S) = softmax(W_s\tilde{h_i}) \tag{4}$$

$\tilde{h_i}$ は文脈ベクトル c_i および隠れ状態の結合により表される．

$$\tilde{h_i} = tanh(W_c[c_i; h_i]) \tag{5}$$

ここで，文脈ベクトル c_i は以下の式から算出される．

$$c_i = \sum_{j=1}^{m} \bar{h}_j a_i \quad (6)$$

上述の式において，$\bar{h}_j$ はステップ j における符号化器の出力である．また，a_i はアライメントベクターと呼ばれ，以下のように表される．

$$a_i = softmax(h_i^\top W_a \bar{h_s}) \quad (7)$$

$\bar{h_s}$ は符号化器から出力される全ての隠れ状態である．注意機構はこれらの式を用いることで，翻訳元文および翻訳先文のアライメントを学習することができる．

4. サブワードユニット

ニューラル機械翻訳は大量のコーパスから長時間の学習を行う必要がある．学習時間を削減するために，低頻度な単語を記号に置き換える手法は文を曖昧にする問題が指摘されており，翻訳精度の低下にも影響している．

この問題を解決するために，近年サブワードユニットを用いた手法 [13] が注目されている．サブワードユニットを生成する手法として，データ圧縮法の一つであるバイト対符号化 [5] が用いられている．バイト対符号化は，学習データに含まれる文字から語彙を作成するため、語彙外語は現れない．しかし，サブワードユニットは**表 1** に示すように文を意味を持たない記号列へと変換する．表 1 において，形態素解析による各トークンは単語の意味を保持している．一方で，バイト対符号化によるトークン列では，“望遠鏡”や “少女”などの単語は分解され意味のないトークンに変換されている．

サブワードユニットを生成する手法の一つである Word-piece モデルは，空白を特殊なメタ記号に置き換えることで，トークン列からの復元を容易にしている [15]．

日本語や中国語，韓国語などのアジア言語では単語間の区切りが存在しない．そのためこれらの言語では，文を単語へ分割するために形態素解析を用いている．形態素解析は辞書に依存しているため，辞書に存在しない固有表現やオノマトペ，他言語で書かれた単語などが出現すると誤りが生じてしまう．また，これを回避するために Web ページなど用いた複雑な手法で辞書を更新する必要ある．つまり，翻訳精度は形態素解析器の性能に依存していると言える．サブワードユニットでは，文を文字列にまで分割することで，言語に依存することなく単語への分割を行うことができる．したがって，サブワードユニットはこれらの課題に対して頑健であると言える．

5. サブワードユニットへの言語学的情報の付与

5.1 直接接続モデル

我々のベースラインモデルを**図 1** に示す．ベースラインモデルには 3.1 節および 3.2 節で示した系列変換モデル及び注意機構を実装した．また，注意機構には直前のステップの出力を入力として用いる input-feeding アプローチ [9] を，再帰型ニューラルネットワークユニットには LSTM を採用している．

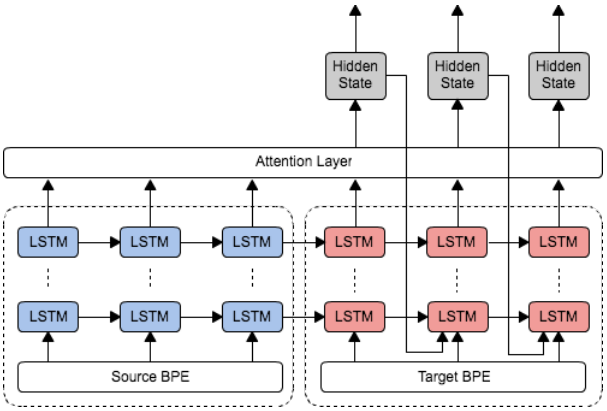

図 1 ベースラインモデル
Fig. 1 Baseline Model

表 1 で示したように，サブワードユニットは意味を持たない場合がある．そのため，文が持つ言語学的な情報は消失してしまう．そこで我々は，ニューラル機械翻訳において言語学的な情報は翻訳精度に寄与するか，新たなモデルを用いて実験を行う．

形態素解析によって得られる形態素情報（表層形，品詞，活用形）は言語学的に正しいと言える．そのため我々は，形態素情報をサブワードユニットに付与することで翻訳精度の向上を図る．しかしながら，サブワードユニットに対して形態素情報を直接付与すれば，語彙数が増大することが考えられる．そのため今回の実験では，形態素情報をサブワードユニットに直接付与するのではなく，隠れ層に変換して用いる．

我々の提案モデルを**図 2** に示す．我々のモデルは，サブワードユニット符号化器および形態素情報符号化器の二種類の符号化器を含んでいる．それぞれサブワードユニットおよび形態素情報を隠れ状態へ変換する．そして，これらの二種類の符号化器を直接接続することで，サブワードユニットに対して形態素情報による言語学的な補完を試みる．ここで，アライメントを明確にするため，注意機構にはサブワードユニットの情報のみ用いる．

我々はサブワードユニットおよび形態素情報間の距離を最小限にすべきだと考えている．これは距離が遠すぎることにより，文脈を保持する勾配が消失してしまうのを防ぐためである．また，本実験のモデルはサブワードユニットおよび形態素情報の可変長の入力が可能であるため，様々な形態素情報を適用することができる．以上の理由から我々は二種類の符号化器を直接的に接続するモデルを提案した．

表 1　バイト対符号化を用いた日本語文のトークン化例
Table 1　Tokenization example using byte pair encoding on Japanse

原文	私は望遠鏡を持った少女を見た.
形態素解析	私 / は / 望遠鏡 / を / 持っ / た / 少女 / を / 見 / た / .
バイト対符号化	私 / は / 望 / 遠 / 鏡 / を持った / 少 / 女 / を見 / た / .

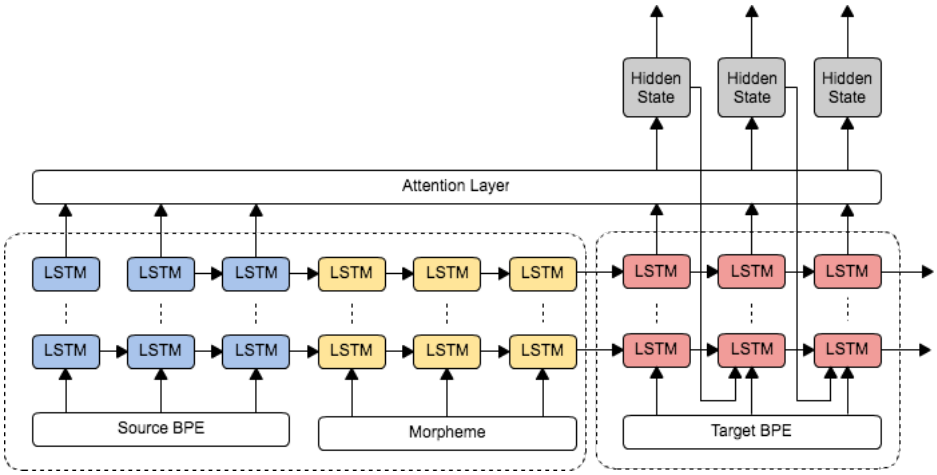

図 2　直接接続モデル
Fig. 2　Directly Connection Model

5.2　形態素情報

本実験では，まず形態素情報として形態素解析により分割された単語列を用いる．言い換えれば，形態素情報として表層形を用いる．

しかしながら表層形を用いる場合，サブワードユニットと入力長が異なるため，入力間のアライメントを評価することは難しい．そのため，我々はサブワードユニットと入力長が同一の形態素情報を提案する．**図 2** に示す 6 つのタグを新しく定義し，これらを用いたタグ付け例を**図 3** に示す．

タグ	意味
B	形態素の始端を示す
M	形態素の中間を示す
E	形態素の終端を示す
W	形態素と一致する
C	複数の形態素から成る
H	メタ文字を示す

表 2　形態学的タグ
Table 2　Morphological tags

6.　実験

今回の実験で用いたモデルはニューラルネットワーク用ライブラリの Chainer[*1] により実装を行った．

[*1] https://chainer.org/

6.1　データセット

日英翻訳を対象に実験を行い，データセットには ASPEC を用いた [10]．学習データは ASPEC に含まれる学習データ *Train-1* の 100 万文対を用いた．同様に ASPEC のテストデータ 1,812 文対を用いてテストを行った．語彙数はサブワードユニットおよび表層形ともに 8,000 に設定し，学習データから語彙を作成した．各データセットにおける全トークン数は**図 4** に示す通りである．また，表層形を用いる場合，頻度が上位 8,000 を下回る単語は<unk>記号に置換した．形態素解析には形態素解析器である MeCab[*2] を用いた．さらにサブワードユニットの生成には Sentencepiece[*3] に実装されているバイト対符号化手法を用いた．

6.2　学習設定

本実験に用いるモデルの LSTM 層数は 1 とした．[*4] 単語分散表現の次元数および隠れ状態の次元数は 1,000 とした．またパラメータの更新には Adam アルゴリズム [7] を用いる．またミニバッチ数は 64，ドロップアウトは 0.2 に設定した．5.2 節で述べたように，形態素情報には表層形および形態学的タグを用いた．

6.3　実験結果

BLEU 値 [12] を用いて翻訳精度の評価を行った．ベース

[*2] http://taku910.github.io/mecab/
[*3] https://github.com/google/sentencepiece
[*4] 層数を増加させた場合，性能の向上が考えられるが，計算時間のを考慮して浅い層のモデルを用いた．

表 3　タグ付け例
Table 3　Tagging example

形態素	現在/筋ジストロフィー/患者/の/移動/介助/において/文書/マニュアル/を/使用/し/て/いる/.
サブワードユニット	現在/,/筋/ジストロフィー/患者の/移動/介/助/において/文書/マ/ニュ/アル/を使用/している/.
タグ	W W B E C W B E W W B M E C C W

表 4　トークン数
Table 4　Token size

Corpus		BPE	Morpheme
Japanese	train	26750747	29155391
	test	46317	50310
English	train	32390641	26685582
	test	55182	46010

ラインには図 1 に示す系列変換モデルおよび注意機構を用いた．また形態学的情報の効果を評価するために，図 2 に示すモデルを用いて実験を行った．ここで，LSTM が増加したことによる性能の上昇を考慮し，二種類の符号化器の両方に対してサブワードユニットを適用したモデルについても実験を行った．実験結果を**表 5** に示す．

表 5　実験結果
Table 5　Main result

System	BLEU
ベースライン	12.27
サブワードユニットおよびタグ	12.35
サブワードユニット（複数）	12.47
サブワードユニットおよび表層形	**12.62**

GPU(Tesla P40) を用いた各モデルにおける 1 エポック毎の学習時間は約 4 時間であった．またテスト時における翻訳時間は約 6 分であった．ベースラインモデルによる翻訳文の BLEU 値は 12.27 となった．またベースラインと比較して，複数のサブワードユニットを適用したモデルでは 0.20 ポイントの上昇を得られた．また形態学的タグを用いたモデルでは 0.08 ポイントの上昇であった．さらに表層形を用いたモデルでは 0.35 の上昇を得ることができた．

7. 考察

全ての直接接続モデルの結果がベースラインモデルの性能を上回った．加えて表層形を用いたモデルは複数のサブワードユニットを用いるモデルの性能を上回った．この結果から，言語学的情報としての表層形はサブワードユニットを用いたモデルに対して効果的であることが分かる．

一方でタグを用いたモデルでは，複数のサブワードユニットを用いたモデルを下回る結果となった．下回った理由として，表層形を用いたモデルの語彙数が 8,000 であったのに対して，タグの種類数が 6 であったため情報量が落ちたことが考えられる．しかしながら，サブワードユニットと入力長を一致させることができた．そのため今後，双方向再帰型ニューラルネットワークなどを用いる様々なモデルに適用できることが考えられる．

7.1　入力長による影響

我々のモデルにおいて，言語学的情報は入力長を長くし，サブワードユニット符号化器と復号化器間の距離を広げる．我々はこの間隔が BLEU 値に影響を及ぼしていると考え，ベースラインモデルおよび表層形を用いた直接接続モデルにおける入力長毎の BLEU 値を算出した．この結果を**図 3** に示す．両モデルとも入力長が増加すると BLEU 値が低下することが分かる．また，入力長が短い文において表層形モデルはベースラインモデルを上回る性能であるが，入力長が 60 を超えるとベースラインを下回っている．この結果から，表層形を用いることで入力列が長くなり、翻訳能力が低下することが分かった．今後これを解決するために，入力長を増加させない別のモデルを考案する必要がある．

7.2　表層形列に含まれる語彙外語

さらに，我々は表層形列に含まれる語彙外語の影響について評価を行った．**表 6** にベースラインから表層形モデルの間で BLEU 値が変化した文についての集計を示す．表における *up* 列はベースラインから性能が上昇した文集合を示しており，同様に *down* 列は性能が低下した文集合を示している．変化のしきい値として 10 以上 BLEU 値が変化した文を対象としている．またどちらの文も性能が高い，もしくは低いものは集計の対象外とした．BLEU 値が低下した文集合に含まれる語彙外語の割合は 4.85% であり，この数値は BLEU 値が上昇した文集合の値よりも高い．したがって，表層形列に多くの語彙外語が含まれる場合，翻訳性能が低下することが分かる．今後の展望として，BLEU スコアを上昇させるために，語彙内の高頻度語へ置換するなど語彙外語を減少させる手法を検討する．

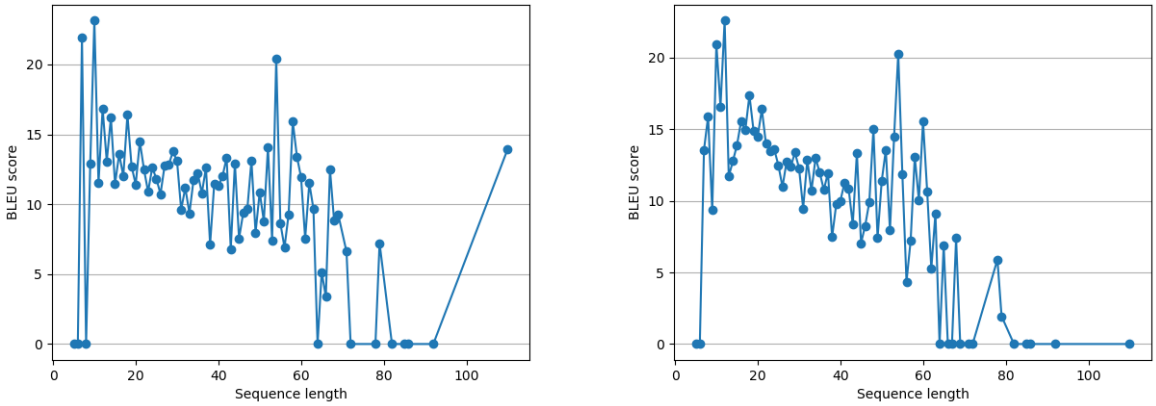

図 3　ベースラインモデル（左）および表層形モデル（右）における入力長毎の BLEU 値
Fig. 3　BLEU score per sentence length. Left:Baseline model. Right:Our proposed model.

表 6　ベースラインから BLEU 値が変化した文についての集計
Table 6　Totalization of changes in the BLEU score from baseline

	up	down	even
sentence	465	435	897
token	12198	11713	25756
unk	523	568	1180
unk / token	0.0429	0.0485	0.0458

7.3　翻訳例

表層形モデルにおける翻訳例を**表 7** に示す．これらは ASPEC のテストデータに含まれる文である．翻訳先ラベルは人手による翻訳である．ベースラインモデルにおいて訳抜けにより誤訳となっていた文について，表層形モデルにおいて言語学的情報がこれらを補完することで正しく翻訳できていることが分かる．例えば，1 つめの例における原因を示す回転速度についての記述や，2 つめの例における実験の詳細についての記述などである．

8.　おわりに

サブワードユニットは低頻度語や未知語問題を解決する非常に有効な手法であるが，文を意味を持たないトークン列へと変換する．そこで本研究では，サブワードユニットに対して言語学的な情報として形態素情報を導入した．その結果，翻訳精度が上昇し，形態素情報はサブワードユニットを用いたモデルに対して有用であることが分かった．一方で，形態素情報によって入力列が長くなるなどの課題が明らかになった．今後の発展として，これらの問題を解決することで，さらなる性能の上昇を試みる．

また，我々のベースラインモデルを深い層で実験を行った．ベースラインの実験設定において LSTM の層数を 4，ドロップアウトを 0.5，ユニット数を 1024 に変化させたモデルを用いた．さらに復号化時にビームサイズ 20 のビームサーチを実装した．このモデルでテストを行ったところ，BLEU 値は 19.72 を記録した．この結果は現在の最先端のモデルの結果に到達していない．そのため，ベースラインの性能を上昇させる必要があることも分かった．

参考文献

[1] Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. and Bengio, Y.: Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1724–1734 (2014).
[2] Costa-jussà, M. R. and Fonollosa, J. A. R.: Character-based Neural Machine Translation, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 357–361 (2016).
[3] Dzmitry, B., Cho, K. and Bengio, Y.: Neural Machine Translation by Jointly Learning to Align and Translate, *ICLR 2015* (2015).
[4] Eriguchi, A., Hashimoto, K. and Tsuruoka, Y.: Tree-to-Sequence Attentional Neural Machine Translation, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 823–833 (2016).
[5] Gage, P.: A New Algorithm for Data Compression, *C Users J.*, pp. 23–38 (1994).
[6] Hochreiter, S. and Schmidhuber, J.: Long Short-Term Memory, *Neural Computaion 9.*, pp. 1735–1780 (1997).
[7] Kingma, D. P. and Ba, J.: Adam:A Method for Stochastic Optimization, *3rd International Conference for Learning Representations* (2015).
[8] Li, X., Zhang, J. and Zong, C.: Towards Zero Unknown Word in Neural Machine Translation, *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pp. 2852–2858 (2016).
[9] 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*, pp. 1412–1421

表 7　翻訳例
Table 7　Translation Example

翻訳元	しかし，回転速度が大きすぎると，逆向きの変形が生じる．
翻訳先	The too high rotation speed produces the reverse deformation.
ベースライン	However, the deformation of the inverse direction occurs due to the rotation speed.
表層形モデル	However, the deformation of the inverse direction occurs when the rotational speed is too big.
翻訳元	高解像度映像の提示に関する評価実験を行った．
翻訳先	The evaluation experiment on the proposal for high resolution images was carried out.
ベースライン	The evaluation of the resolution image was carried out.
表層形モデル	The evaluation experiment on the presentation of high resolution image was carried out.

(2015).
[10] Nakazawa, T., Yaguchi, M., Uchimoto, K., Utiyama, M., Sumita, E., Kurohashi, S. and Isahara, H.: ASPEC: Asian Scientific Paper Excerpt Corpus, *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (Chair), N. C. C., Choukri, K., Declerck, T., Grobelnik, M., Maegaard, B., Mariani, J., Moreno, A., Odijk, J. and Piperidis, S., eds.), pp. 2204–2208 (2016).
[11] Niehues, J., Cho, E., Ha, T.-L. and Waibel, A.: Pre-Translation for Neural Machine Translation, *Proceedings of the 26th International Conference on Computational Linguistics*, pp. 1828–1836 (2016).
[12] Papineni, K., Rukos, S., Ward, T. and Zhu, W.-J.: BLEU: A Method for Automatic Evaluation of Machine Translation, *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pp. 311–318 (2002).
[13] 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*, pp. 1715–1725 (2016).
[14] Sutskever, I., Vinyals, O. and Le, Q. V.: Sequence to Sequence Learning with Neural Networks, *Proceedings of the 27th International Conference on Neural Information Processing Systems*, pp. 3104–3112 (2014).
[15] Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., ukasz Kaiser, Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M. and Dean, J.: Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation, *CoRR* (2016).
[16] Zhang, J., Li, L., Way, A. and Liu, Q.: Topic-Informed Neural Machine Translation, *Proceedings of the 26th International Conference on Computational Linguistics*, pp. 1807–1817 (2016).