固有表現情報を用いたニューラル機械翻訳

鵜川 新[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 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
[3] 東京工業大学 未来産業技術研究所，[4] 産業技術総合研究所
{ugawa@ai, tamura@, ninomiya@}cs.ehime-u.ac.jp
{takamura, oku}@pi.titech.ac.jp

1 はじめに

近年，自然言語処理の様々なタスクでニューラルネットワーク (NN) が活用され，その有効性が示されている．機械翻訳においても，ニューラルネットワークによる機械翻訳 (NMT) が盛んに研究されている．現在の NMT は，入力系列を中間表現にエンコードする NN とエンコードされた中間表現から出力系列をデコードする NN で構成されるエンコーダ・デコーダモデル [1] が主流であり，このモデルを改良した様々なモデルが提案されている．

一方，単語の言語学的素性を活用する試みがなされている．Sennrich ら [2] は，見出し語，品詞やサブワードといった言語学的素性を埋め込んだベクトルを単語埋め込みベクトルに連結したベクトルをエンコーダでエンコードすることで，NMT の性能改善を実現している．

統計的機械翻訳では，言語学的素性として固有表現の情報が性能改善に寄与する場合があることが知られている [3]．

そこで本研究では，固有表現の情報を用いて NMT の性能改善を試みる．具体的には，提案モデルは，エンコーダ部分で，固有表現タグを埋め込んだベクトルと単語埋め込みベクトルの線形和ベクトルから中間表現を生成することで，原言語側の固有表現情報を考慮した翻訳を行う．ASPEC データを用いた英日翻訳の評価実験により，固有表現情報を考慮することで，BLEU が 0.88 ポイント向上することを確認した．

2 従来の NMT

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 [1] は，エンコーダとデコーダ用の再帰型ニューラルネットワーク（RNN）により翻訳を実現する．RNN としては，GRU（Gated Recurrent Unit）や LSTM（Long Short Term Memory）[4] 等が使われるが，本研究では，LSTM を使用する．

エンコーダ部は，原言語の単語列，$x = (x_1, x_2, ..., x_m)$ を入力し，LSTM により中間状態 $c = (h_1, h_2, \ldots, h_m)$ を生成する．デコーダ部は，生成した中間状態 c から目的言語の単語列 $y = (y_1, y_2, \ldots, y_n)$ を逐次的に出力する．中間状態 h_j は，一時刻前の状態 h_{j-1} と現時刻の入力 x_j から LSTM により，式 (1) で算出される：

$$h_j = f_{enc}(h_{j-1}, E_x(x_j)). \tag{1}$$

ここで f_{enc} は，エンコーダ側の LSTM であり，E_x は，単語埋め込み層とする．

デコーダ部では，出力単語列 y を入力単語列 x に対する対数尤度に基づき算出する：

$$\log p(y|x) = \sum_{i=1}^{n} \log p(y_i|y_{1:i-1}, h_m). \tag{2}$$

各単語 y_i の確率分布は，デコーダにおける中間状態 h_i から計算される：

$$p(y_i|y_{1:i-1}, h_m) = softmax(proj(h_i, c)). \tag{3}$$

$proj$ は，LSTM の中間状態のベクトルを次元が語彙数のベクトルにマッピングを行う関数である．

2.2 アテンション機構

アテンション機構とは，デコード時に入力単語列と出力単語列間で関係性が高い単語対を捉えて利用する機構である．アテンション機構では入力単語ベクトルと出力単語ベクトルとの類似度に基づいたスコア a_{ij}

を算出し，スコア a_{ij} を重みとした中間状態から文脈ベクトル s_i を算出する：

$$a_{ij} = \frac{exp(h_i \cdot h_j)}{\sum_{k=1}^{T_x} exp(h_i \cdot h_k)}, \quad (4)$$

$$s_i = \sum_{j=1}^{T_x} a_{ij} h_j. \quad (5)$$

ここで T_x は，原言語の文長である．デコーダでは，中間状態と文脈ベクトル s_i を用いて，式 (6) により出力単語の確率分布を算出する：

$$p(y_i|y_{1:i-1}, c) = softmax(proj([h_i; s_i])). \quad (6)$$

3 提案手法

本節では，まず，3.1 節で本研究で用いる固有表現を説明し，3.2 節で固有表現情報を用いた NMT を提案する．

3.1 固有表現

固有表現とは，日付，数値，人名，地名などの特定の表現のことであり，単語の情報がどのようなものであるかを識別する指標にすることができる．例えば，‘東京駅に 10 時 20 分に着いた’ は，東京駅は「地名」，10 時 20 分は「時刻」を表す固有表現である．MUC（Message Understanding Conference）では，「人名」，「地名」，「組織名」，「時間」，「日時」，「金額表現」，「割合表現」の 7 種類の固有表現を用いている．また，IREX（Information Retrieval and Extraction Exercise）では，上記固有表現に「固有物名」を加えた 8 種類の固有表現を用いている．一般的に用いられている CoNLL2003 データセットでは，「人名」，「組織名」，「地名」，「その他」の 4 種類の固有表現が付与されている．

固有表現の情報は，例えば，IOB2 タグで付与される．IOB2 タグでは，B（Begin），I（Inside），O（Outside）の三種類のを用いて，B と I で固有表現を，O で固有表現以外を表す．例えば，‘東京駅に 10 時 20 分に着いた’ の文には，‘東京:B-地名，駅:I-地名，に:O，10:B-時間，時:I-時間，20:I-時間，分:I-時間，に:O，着い:O，た:O’ のように固有表現の情報を付加する．

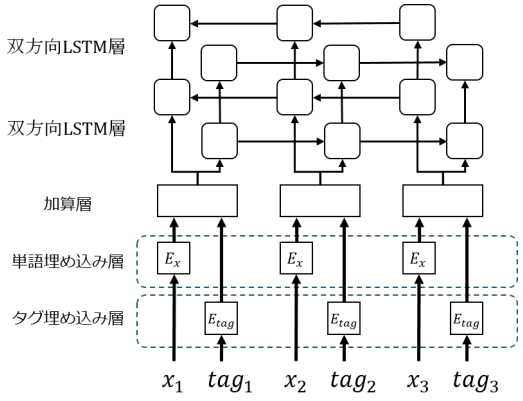

図 1: 固有表現情報を用いたエンコーダ部

3.2 固有表現情報を用いた NMT

本研究では，原言語の固有表現情報を活用する NMT を提案する．具体的には，固有表現タガーを用いて原言語の文に固有表現タグを付与し，原言語の文と共に固有表現タグをエンコーダの入力とする．これにより，原言語中の固有表現に対する翻訳精度の向上を狙いとしている．

提案手法では，データスパースネスの問題を避けるために，固有表現の情報として，IOB2 タグの B タグと I タグを区別せずにラベルクラス情報と O タグの情報を用いる．例えば，‘東京駅に 10 時 20 分に着いた’ の文に対しては，‘東京:地名，駅:地名，に:O，10:時間，時:時間，20:時間，分:時間，に:O，着い:O，た:O’ のように固有表現の情報を付加し，NMT の入力とする．図 1 に提案モデルにおけるエンコーダ部を示す．提案モデルでは，単語埋め込み層 E_x で原言語の単語列に対する埋め込みベクトルを生成する．加えて，タグ埋め込み層 E_{tag} を用意し，固有表現タグに対する埋め込みベクトルを生成する．そして，加算層で両埋め込みベクトルの線形和ベクトルを算出し，LSTM エンコーダの入力とする．したがって，エンコーダの中間状態は，式 (7) のように算出される：

$$h_j = f_{enc}(h_{j-1}, (E_x(x_j) + E_{tag}(tag_j) * bias)). \quad (7)$$

式 (7) において，tag_j は単語 x_j に対して付与された固有表現タグを表し，固有表現情報の影響度を $bias$ により指定している．

4 実験

4.1 対訳データセット

データセットには，Work shop on Asian Translation（WAT）で用いられた Asian Scientific Paper Excerpt Corpus（ASPEC）[5] の英日対訳コーパスを利用した．日本語は，KyTea[6] を用いて単語区切りを行った．英語は，spaCy[1]を用いてトークン化した．翻訳モデルの学習には日本語，英語ともに 50 単語未満の対訳文を用いた．また，頻度が 2 回以上の単語のみを語彙として用い，頻度 1 回の単語は ⟨UNK⟩ タグに置き換えた．学習データは，20,000 文対，100,000 文対を用いた 2 種類の翻訳性能を評価する．学習データ 20,000 文対の語彙サイズは，9,497(英語)，8,712(日本語)，学習データ 100,000 文対の語彙サイズは，24,923(英語)，23,304(日本語) であった．開発データは，1768 文対，テストデータは，1755 文対を用いた．

4.2 固有表現情報の付与

固有表現情報は，OntoNotes 5.0[2]と Common Crawl[3] から学習した NamedEntityRecognition101[1][4] により付与した．使用した固有表現タグを表 1 に示す．また，英日コーパスの原言語側（英語）に付与された固有表現タグの割合を表 2 に示す．

4.3 比較手法

本研究では，英日翻訳を通じて提案手法の有効性を検証する．ベースラインとして，エンコーダ部を双方向 LSTM2 層，デコーダ部を LSTM2 層で構成したアテンション付きエンコーダ・デコーダモデルを使用する．提案手法のエンコーダ部も双方向 LSTM2 層を用いたが，エンコーダ部の入力は，単語埋め込みベクトルと固有表現タグの埋め込みベクトルを足し合わせたベクトルであることを特筆しておく．デコーダ部は，ベースライン同様，LSTM2 層で構成したアテンション付きエンコーダ・デコーダモデルを使用する．*bias* は 0.5 とした．提案手法，ベースラインともに，エンコーダ，デコーダの全埋め込み層（単語埋め込み層，固有表現タグ埋め込み層）のサイズは 256 次元，LSTM 隠れ層のサイズは 256 次元とした．

[1]https://spacy.io/
[2]https://catalog.ldc.upenn.edu/ldc2013t19
[3]http://commoncrawl.org/
[1]https://spacy.io/usage/linguistic-features#101

表 1: 固有表現タグの種類

固有表現タグ	詳細
PERSON	仮想的な人物を含む人々など
NORP	国籍，宗教や思想など
FACILITY	空港や高速道路，橋などの建物
ORG	会社，店など
GPE	国，都市など
LOC	山の高さや海の深さなど
PRODUCT	車や食べ物などのモノ
EVENT	スポーツなどの大会や行事
WORK_OF_ART	小説や歌など
LAW	法律に関すること
LANGUAGE	言語
DATE	日付
TIME	日付より小さい単位の時間
PERCENT	割合 (%を含む)
MONEY	金額 (単位を含む)
QUANTITY	重さや長さの量
ORDINAL	順序
CARDINAL	他に該当しない数値
OTHER	上記に該当しなかったもの

表 2: コーパスに付与された固有表現タグの割合

対訳文数	情報の上位 3 つ
20,000 文	O:95.3%，ORG:1.74%，DATE:0.85%
100,000 文	O:95.2%，ORG:1.62%，DATE:0.9%

4.4 実験結果

翻訳性能は，BLEU[7] で評価した．評価結果を表 3 に示す．表 3 より，学習データ 20,000 文対では 0.58 ポイント，100,000 文対では 0.88 ポイント，提案手法はベースラインの性能を上回った．

5 考察

提案手法は，ベースラインと比較してエンコーダ部に固有表現情報を用いることで，BLEU 値が向上して

表 3: ベースラインと提案手法の BLEU 値

	対訳文数	
	20,000	100,000
ベースライン	15.29	24.70
提案手法 bias 値 0.5	15.87	25.58

表 4: 各手法の翻訳例

入力文	about	60	%	of	the	first	peak	showed	ca2	+	dependence	.
入力情報	PERCENT	PERCENT	PERCENT	O	O	ORDINAL	O	O	O	O	O	O
参照訳: 最初 の ピーク の 約 6 割 が c a 2 + 依存 性 を 示 し た .												
ベースライン: 1 つ の ピーク の 約 6 0 % は c a + 依存 性 を 示 し た .												
提案手法: 最初 の ピーク の 約 6 0 % は c a 2 + 依存 性 を 示 し た .												

いる．しかし，固有表現情報の‘O’は，全体の約 95%を占めているため，その他の固有表現情報に効果があると考えられる．そのため，‘O’以外の固有表現情報が多く含まれるデータに対しては提案手法がより効果を発揮すると期待される．

表 4 にベースラインと提案手法が翻訳した実例を示す．入力文には，固有表現情報として PERCENT と ORDINAL が付与されている．この例において，ベースラインは‘first’を‘1 つ’と誤って訳しているのに対し，提案手法は‘最初’と正しく翻訳できている．これは‘first’の固有表現 ORDINAL を活用できたためと考えられる．

今回，提案手法のモデルの固有表現情報の影響度 bias を 0.5 に設定したが，bias を 1.0 にすると BLEU 値が減少した．このことから，固有表現を有効活用するには bias を調整する必要があると考えられる．現在は固有表現情報に統一した bias を用いているが，今後，固有表現タグごとに bias を変更することでさらなる精度向上が期待できる．

6 おわりに

本研究では，固有表現情報を用いるニューラル機械翻訳モデルを提案した．今回の研究では，‘O’以外のタグ数が少ないため IOB2 タグの I タグと B タグを区別していない．今後は，データ数を増やし，O 以外のタグの数を増加させ，固有表現情報の有効性を確かめたい．

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] 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] Rico Sennrich and Barry Haddow. Linguistic input features improve neural machine translation. In *WMT*, 2016.

[3] Santanu Kumar Pal, Sudip Kumar Naskar, Pavel Pecina, Sivaji Bandyopadhyay, and Andy Way. Handling named entities and compound verbs in phrase-based statistical machine translation. 2010.

[4] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. Vol. 9, pp. 1735–1780. MIT Press, 1997.

[5] 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 10th Conference on International Language Resources and Evaluation (LREC2016)*, 5 2016.

[6] Graham Neubig, Yosuke Nakata, and Shinsuke Mori. Pointwise prediction for robust, adaptable japanese morphological analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pp. 529–533, 2011.

[7] Kishore Papineni, Salam 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. 311–318, 2002.