ニューラル機械翻訳における長文分割によるコーパスの拡張

張 津一　　松本 忠博
岐阜大学 大学院 工学研究科
{zhang, tad}@mat.info.gifu-u.ac.jp

1 はじめに

近年，ニューラル機械翻訳（Neural Machine Translation， NMT）の登場によって流暢で精度の高い翻訳が可能となってきた．NMT では翻訳の品質が学習のための対訳データの量に強く依存し，質の高い翻訳結果を得るには大量の対訳データを必要とする．しかし，英語を含む言語対や欧州の言語間の言語対などを除き，一般に十分な量の対訳データを入手するのは困難である．これは NMT の大きな問題点の一つであり，その解決策がいくつか提案されてきた．

転移学習を利用して，十分な量の対訳データを持つ言語対の NMT モデルのパラメータを少ない量の対訳データを持つ言語対の NMT モデルへ転移するアプローチは，この問題の解決策の一つである．Firat ら [1] はこの考え方に基づいて，訓練された豊富な量の対訳データを持つ言語対（フランス語-英語）NMT モデルを親モデル，少ない量の対訳データを持つ言語対（スペイン語-英語）の NMT モデルを子モデルとし，親モデルのいくつかのパラメータを子モデルに転送することによって，少ない量の対訳データを持つ言語間の翻訳精度を大幅に改善した．しかしこの方法には，親モデルと子モデルは類似の言語構造を持たなければならないという制約がある..

他の解決策としてデータ拡張（水増し）の手法がいくつか提案されている．Fadaee ら [2] は，対訳文中の低頻度語を別の単語に置換して得られた文を学習データに加えることで翻訳性能が向上することを示した．Sennrich ら [3] は，目的言語の単言語コーパスを機械翻訳によって原言語へ逆翻訳することで擬似的な対訳文を生成し，対訳コーパスと混合して訓練する方法を提案している．Currey ら [4] は，目的言語の単言語コーパスの文をそのままコピーして原言語側のデータとして用いるだけでも翻訳精度が向上することを示した．

本論文では，既存の対訳データを元にデータ拡張を行う方法を提案する．NMT では文が長くなると翻訳精度が低下する．とくに文字レベルの学習データは単語レベルのものよりさらに長くなるため，訳抜けや訳語の重複が発生しやすくなる．そこで，単語アラインメント情報を利用して長い対訳文から短い対訳文（対訳部分文）を作成し，目的言語側の短文を NMT で逆翻訳して原言語短文を得た後，元の原言語文の一部を逆翻訳結果の短文と入れ替えて擬似的な原言語文を生成して対訳データを拡張する．

評価実験は日中・中日 NMT により行った．表語文字である漢字を主に使用する中国語では 1〜2 文字の単語が多く，同じく漢字を利用する日本語との間の翻訳では文字レベルの NMT が適していると考えられ，また，文字レベルの NMT では，文を単語に分割する過程での誤りや揺れが生じないという利点もあることから，学習と翻訳は文字レベルで行った．

NMT モデルとして Luong ら [5] のものを，実験用データとしてアジア学術論文抜粋コーパス（ASPEC）[6] を用いて評価実験を行った結果，単純に対訳データをコピーして水増しした場合よりも高い BLUE スコアが得られた．

2 NMT システム

本研究では Luong ら [5] によるグローバル注意機構付きエンコーダ・デコーダモデルを実装した NMT システムを文字レベルで使用する．エンコーダは，双方向 LSTM リカレントニューラルネットワークであり，入力系列 $\boldsymbol{x} = (x_1, \cdots, x_m)$ を読み取って，順方向の隠れ状態列 $(\overrightarrow{h_1}, \cdots, \overrightarrow{h_m})$ と逆方向の隠れ状態列 $(\overleftarrow{h_1}, \cdots, \overleftarrow{h_m})$ を求める．隠れ状態 $\overrightarrow{h_j}$ と $\overleftarrow{h_j}$ は連結され，アノテーションベクトルが作られる．

デコーダは，目的言語文 $\boldsymbol{y} = (y_1, \cdots, y_n)$ を予測する LSTM リカレントニューラルネットワークである．

各単語（文字レベルの場合，各文字）y_i は，リカレント隠れ状態 s_i と，前回予測された単語（または文字）y_{i-1}，文脈ベクトル c_i を元に予測される．文脈ベクトル c_i は，アノテーション h_j の加重和として計算される．各 h_j の重みは，y_i と x_j のアラインメントについての情報を表すモデル α_{ij} を通じて決められる．

エンコーダの順方向状態は以下のように表される．

$$\vec{h_j} = \tanh(\vec{W}Ex_j + \vec{U}\vec{h}_{j-1}) \quad (1)$$

ここで，$E \in \mathbb{R}^{p \times V_x}$ は単語埋め込み行列であり，$W \in \mathbb{R}^{q \times p}$ と $U \in \mathbb{R}^{q \times q}$ は重み行列である．p，q，V_x はそれぞれ，単語ベクトルのサイズ，隠れユニットの数，原言語の語彙サイズである．

NMT システムの実装としては OpenNMT[7] を用いた．

3　長文の分割によるコーパスの拡張

Sennrich ら [8] は既存の対訳データとは別に，目的言語の単言語コーパスを用意し，それを原言語へ逆翻訳することで対訳データを増やす方法を提案した．本研究では図 1 に示すように，対訳データ中の長い文（読点等で区切られた文）を元にデータ拡張を行う．NMT では，文が長くなると翻訳精度が低下するが，単語レベルに比べて文字レベルの学習ではさらに文が長くなる．そこで，対訳データの文から以下のような手順で短い対訳文（部分文）を生成して対訳データを拡張する．

3.1　対訳部分文の生成

以下の手順により，学習データに含まれる対訳文から短い対訳文（部分文）を生成する（図 1 前半）．

1. 対訳文を単語に分割した後，単語アラインメント情報を取得する．単語分割には MeCab[*1]と jieba[*2]を，単語アラインメント情報の取得には fast_align[*3]を用いた．
2. 句読点などの記号（, ， . ． ？ ！ ; ： ’）の箇所で対訳文をセグメントに分割する．図 2 に単語アラインメント情報とセグメント分割の例を示す．
3. 原言語文の各セグメントについて，セグメント内の各単語に対応する単語が目的言語側のどのセグメントに属すかを単語アラインメント情報を元に調べてカウントし，セグメント間の対応の割合を求める（対応関係のない単語は割合の計算から除外する）．その割合がある閾値（後述の実験では 0.5 とした）以上のとき，そのセグメントに対応すると考える．
4. 目的言語の各セグメントについても同様にして，対応する原言語セグメントを求める．
5. セグメント間の対応関係が 1 対多や多対多の場合は，1 対 1 になるように複数側のセグメントを 1 つにまとめる．このようにして対訳部分文を作成する．図 3 の例では，日本語文と中国語文はそれぞれ 3 つのセグメントに分割され，2 つの対訳部分文が作られる．

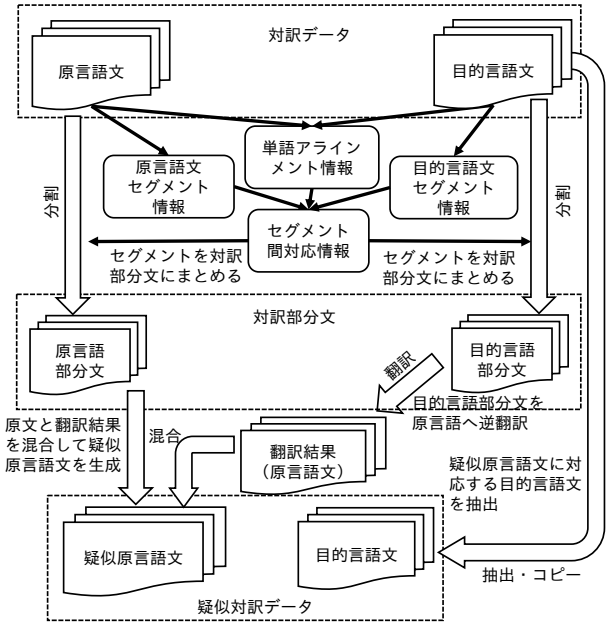

図 1　長文分割によるコーパスの拡張の流れ

この手順で ASPEC-JC の dev データ 2，090 文対から 4，381 文対対訳部分文を生成し，分割位置を人手で確認したところ，概ね 9 割は適切と判断できた．

3.2　対訳データの拡張

生成した対訳部分文を用いて，擬似的な対訳文を以下の手順で構成する（図 1 後半）．なお，対応するセグメントを持たないセグメントが存在する文や，原言語と目的言語でセグメントの順序が異なり，対応が交差

[*1] http://taku910.github.io/mecab/

[*2] https://github.com/fxsjy/jieba（jieba では全角英単語が文字ごとに分割されてしまうため，分割を防ぐ処理を加えた．）

[*3] https://github.com/clab/fast_align

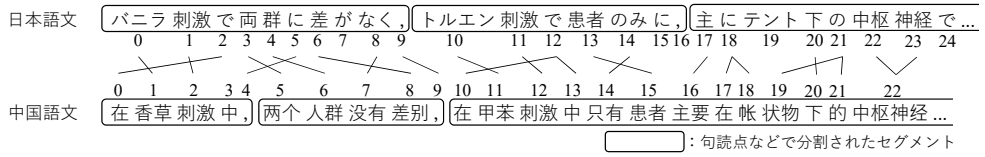

図 2　文のセグメント分割と単語アラインメント情報の例

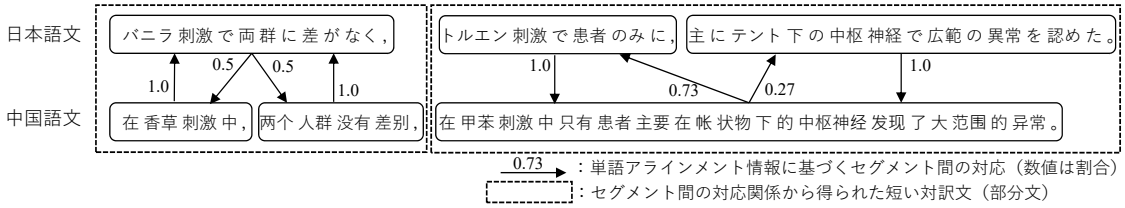

図 3　セグメント間の対応関係と対訳部分文の生成

表 1　生成された擬似原言語文の例（//は分割点）

原文	バニラ刺激で両群に差がなく，//トルエン刺激で患者のみに，主にテント下の中枢神経で広範の異常を認めた．
擬似原言語文 1	香草刺激では，両群に差はなかった // トルエン刺激で患者のみに，主にテント下の中枢神経で広範の異常を認めた．
擬似現言語文 2	バニラ刺激で両群に差がなく，//トルエン刺激には主に帳票物下の中枢神経で広範囲の異常が認められた．

する箇所のある文からは，正しい対訳文が得られない可能性が高いと考え，データ拡張には使用しない．

1. 生成した目的言語の部分文を NMT により原言語に逆翻訳する．逆翻訳のためのモデルは，拡張前の対訳データにより構築したものを使用する．
2. 対訳データの各原言語文の一部を，逆翻訳によって得られた部分文で置き換えて，元の原言語文と一部異なる擬似原言語文を作る．これにより文の分割数と同じ数の擬似原言語文のバリエーションが生成できる．図 3 の日本語文から生成された擬似原言語文を表 1 に示す．
3. 作成された擬似原言語文に対応する目的言語文を用意（コピー）して，擬似的な対訳文とする．

以上の手順で生成した疑似対訳文対を元の対訳データに加えることでコーパスを拡張する．

4　翻訳実験

4.1　NMT システムの設定と翻訳結果の評価法

翻訳システムの実装には OpenNMT を用いた．モデルのパラメータは $(-0.1, 0.1)$ の範囲の一様乱数で初期化し，最適化にはデフォルトの確率的勾配降下法を用いた．学習率はエポック 6 までは 1.0 とし，それ以降はエポックごとに 0.5 倍する．最大勾配ノルムは 1，最大バッチサイズは 100 とした．また，LSTM リカレント層は 1 層で，単語ベクトルと隠れ層の次元は 512 とした．dropout 確率は 0.5 に設定し，デコード時のビームサイズは 5 とした．文の最大長は，デフォルトでは 250 だが，文字レベルでは長くなるため 500 に設定した．

学習と翻訳は文字レベルで行うが，評価は単語レベルのシステムと同じ条件で行うため，日本語文は MeCab，中国語文は jieba で単語ごとに分割した後，OpenNMT 付属の multi-bleu.perl で BLEU スコアを算出した．

多くの場合，エポック 10 前後で validation perplexity（dev データでの perplexity）が下げ止まった．その時点からエポック 16 までの BLEU スコアの平均を評価値とした．ベースラインは何も加工しない訓練データによる文字レベルの翻訳である．

4.2　長文分解による学習データの拡張

対訳文の分割・逆翻訳により拡張した対訳データを用いた翻訳の結果を表 2 に示す．ベースラインは拡張前の訓練データを用いた文字レベル翻訳の結果である．「重複データ」は，比較のために，データ拡張によって

表 2　翻訳実験結果

手法	対訳文数	日本語 → 中国語			中国語 → 日本語		
		語彙サイズ	BLUE (%)		語彙サイズ	BLUE (%)	
ベースライン	672,304	6,082	39.92		4,249	40.52	
重複データ（比較用）	2,122,066	6,082	40.27	(+0.35)	4,249	41.56	(+1.04)
データ拡張（提案手法）	2,122,066	6,082	40.65	(+0.73)	4,249	42.34	(+1.82)

追加された文の原文を単純にコピーして作成したデータである．擬似原言語文には，逆翻訳時の誤りやつながりの不自然な箇所が見られるものの，単純に原文をコピーして水増しした場合と比べて，BLEU スコアは日 → 中で 0.38%，中 → 日で 0.78% 向上した．

5　おわりに

本論文では，対訳コーパス中の句読点等を含む比較的長い文を分割・逆翻訳して，原文の組み合わせて対訳データを拡張する手法を提案した．ASPEC-JC コーパスを用いた文字レベルの日中・中日 NMT により評価実験を行ったところ，BLEU 値は拡張前に比べて，日→中で 0.73%，中→日で 1.82% 向上した．本手法は言語に依存しないため，どの言語対に対しても効果が期待できる．

謝辞

著者の一人である張津一は中国国家留学基金委の助成 (No.201708050078) を受けている．ここで感謝の意を表する．

参考文献

[1] O. Firat, K. Cho, and Y. Bengio, “Multi-Way, Multilingual Neural Machine Translation with a Shared Attention Mechanism,” ArXiv e-prints, pp.866–875, 2016. http://arxiv.org/abs/1601.01073

[2] M. Fadaee, A. Bisazza, and C. Monz, “Data augmentation for low-resource neural machine translation,” Proc. 55th Annual Meeting of the Assoc. for Computational Linguistics (Volume 2: Short Papers), pp.567–573, Vancouver, Canada, 2017.

[3] R. Sennrich, B. Haddow, and A. Birch, “Improving neural machine translation models with monolingual data,” Proc. 54th Annual Meeting of the Assoc. for Computational Linguistics, pp.86–96, Belin, Germany, Aug. 2016.

[4] A. Currey, A.V.M. Barno, and K. Heafield, “Copied monolingual data improves low-resource neural machine translation,” Proc. Conf. on Machine Translation (WMT), pp.148–156, Copenhagen, Denmark, Sept. 2017.

[5] M.T. Luong, H. Pham, and C.D. Manning, “Effective approaches to attention-based neural machine translation,” Proc. 2015 Conf. on Empirical Methods in Natural Language Processing, pp.1412–1421, ACL, Lisbon, Portugal, 2015.

[6] T. Nakazawa, M. Yaguchi, K. Uchimoto, M. Utiyama, E. Sumita, S. Kurohashi, and H. Isahara, “ASPEC: Asian scientific paper excerpt corpus,” Proc. 9th Int. Conf. on Language Resources and Evaluation (LREC 2016), pp.2204–2208, European Language Resources Association (ELRA), Portorož, Slovenia, 2016.

[7] G. Klein, Y. Kim, Y. Deng, J. Senellart, and A.M. Rush, “OpenNMT: Open-source toolkit for neural machine translation,” ArXiv e-prints, p.1, 2017.

[8] R. Sennrich and B. Haddow, “Linguistic input features improve neural machine translation,” Proc. 1st Conf. on Machine Translation, vol.1, pp.83–91, 2016.