訳抜けを含む訓練データと訳抜けのない出力とのギャップを埋めるニューラル機械翻訳

後藤 功雄　　美野 秀弥　　山田 一郎

NHK 放送技術研究所

{goto.i-es, mino.h-gq, yamada.i-hy}@nhk.or.jp

1 はじめに

機械翻訳では，入力された文の内容を全て翻訳することが望ましい．このような機械翻訳システムを構築するためには，対訳関係に過不足のないデータで学習することが理想的である．しかしながら，既存の対訳データには，対訳関係にノイズが多い場合がある．国内向けのニュース記事を元に，英語ニュースを制作する場合，単なる翻訳ではなくニュースライティングとなり，きれいな対訳関係にならない [8]．例えば，日本語ニュースでは記事中に重複した内容がしばしば現れるが，英語では記事中での内容の重複は避ける傾向があるため，重複部分は省く．外国向けの英語ニュース記事では，国内向けのニュース記事から，外国人には重要ではない詳細な話を省く．このようにして，対訳関係において，目的言語側で内容の不足が発生する[*1]．図 1 に，日本語ニュース (a) に対して英語ニュース (b) の内容に不足がある対訳文の例を示す．日本語文で赤文字の内容が英語文に含まれていない．このように原言語文の内容が目的言語文に含まれていない場合を「訳抜け」と呼ぶ．機械翻訳は (a) を入力したら (b) を出力するように学習する．すなわち訳抜けも学習してしまうので，翻訳時も (a) を入力すると (b) が出力されやすくなる．一方で，入力文を正確に翻訳することを目標とした機械翻訳では，(a) を入力すると (c) を出力してほしい．この (b) と (c) とのギャップが訳抜けのない出力の実現において課題である．訳抜けを多く含む訓練データで学習しても，入力文の翻訳では訳抜けしない翻訳を実現したい．

本稿ではこの課題を解決するため，この訓練時と推論時とのギャップを埋めるニューラル機械翻訳（NMT）手法を提案する．提案手法は，訓練データ中において，目的言語文に訳出されていない内容と目的言語文に訳出されている内容を区別するラベルを原言語文中の各単語に対応させて生成する．このラベル系列を使うことで，NMT は訓練データで訳出されていない部分と訳出されている部分を区別して学習できる．推論時には，訓練データで訳出されている部分から学習した知識を主に使って入力文を翻訳する．報道記事の日英翻訳の実験を実施して，ベースラインの Transformer[12] と比較して，提案手法は訳抜けが 24.8% 減少したことを確認した．

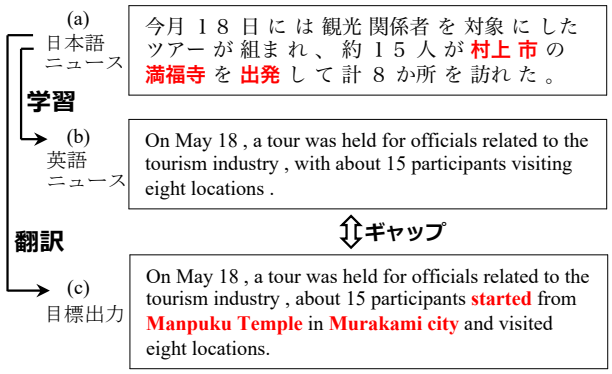

図 1　翻訳不足のあるニュース対訳文の例（読売新聞/Daily Yomiuri）．日本語の赤色部分の内容が英語ニュースに含まれていない．機械翻訳は (a) から (b) への翻訳を学習するが，翻訳時は (a) から (c) の出力が目標となる．

2 提案手法

2.1 訳抜けの有無を区別した学習によるギャップの橋渡し

訓練時に NMT が図 1(a) から (b) を学習した場合，推論時に NMT は入力が (a) なら (b) を出力しやすくなる．しかしながら，NMT の目標出力は (c) である．この (b) と (c) のギャップを橋渡しするため，訓練データで訳出されていない部分と訳出されている部分を区別するラベルを使って，これらを区別して NMT モデルを学習する．訓練データの原言語の内容語で訳出されていない語にラベル D (Deficient)，それ以外の原言語の語にラベル t (translated) を付与する（図 2(a)）．推論時，入力文の語にラベル t を付与することで，NMT は訳出されている部分から学習した知識を主に用いて翻訳できる（図

[*1] 不足だけでなく，情報の追加もある．例えば地名に加えて大まかな場所（北日本）や地震の震度に加えて日本の震度はスケールが 7 段階であるなどがある．情報の追加は本稿では扱わない．

ラベル: t t t t D D t D t D t t t
原文: 約 1 5 人 が **村上 市** の **満福寺** を **出発** し た 。

(a) 訓練データのラベル

ラベル: t t t t t t t t t
原文: 村上 市 で 花火 大会 が あり ま し た 。

(b) テストデータのラベル

図 2　ラベルの例．赤色の単語の内容は目的言語文に含まれていないことを想定している．

2(b))．

NMT では，各原言語単語に対応したラベルをラベルの分散表現に変換し，対応する原言語単語の分散表現とラベルの分散表現との和を計算する．この分散表現の和を従来の NMT の原言語単語の分散表現と同様に扱う．以下，ラベルの自動付与手法を説明する．

2.2　訓練データへのラベルの付与

訓練データへラベルを付与するためには，訳抜けの検出が必要である．この検出には，単語対応確率 [9]，および逆翻訳確率 [3] が手がかりとして利用できる．本稿では，これらを実験で比較する．

■単語対応を検出に用いる方法　対訳文において，訳抜けがなければ，多くの場合に原言語の内容語は目的言語の内容語に対応すると考えられる．

そこで，対訳文の単語対応を計算し，内容語と対応していない原言語の内容語を訳抜けしていると判別する．ここでは，後の実験で用いる [9] の単語対応確率を使った内容語の訳抜け判別方法を利用する．[9] は，RNN ベースの NMT のアテンション確率の計算において，出力単語の単語分散表現もアテンションション確率の計算に利用する．これよって，次に翻訳すべき箇所の予測ではなく，出力した単語に対応する箇所の推測となる．本稿ではこの確率をアラインメント確率と呼ぶ．原言語文を $\mathbf{x} = x_1, ..., x_{Tx}$，目的言語文を $\mathbf{y} = y_1, ..., y_{Ty}$ とし，目的言語単語 y_i が原言語単語 x_j に対応するアラインメント確率を $p(y_i \rightarrow x_j)$ と表す．双方向のアラインメント確率において，$\max_{i|y_i \in C} p(y_i \rightarrow x_j)p(x_j \rightarrow y_i)$ が閾値未満の内容語 x_j が訳抜けしていると判別する．ここで，C は内容語の集合を表す．この方法を NMT alignment と呼ぶ．

■逆翻訳確率を検出に用いる方法　入力文中に存在しない内容語の訳語は，NMT での目的言語から原言語への逆翻訳確率が低くなると期待される．そこで，目的言語文を原言語文に翻訳する逆方向の NMT を学習する．対訳文対の目的言語文を NMT に入力して，対訳文対の原言語文を出力する時の各単語の生成確率が，閾値未満の場合に訳抜けしていると判別する．この方法を BT probability と呼ぶ．

■単語対応と逆翻訳確率の両方を検出に用いる方法　単語対応と逆翻訳確率には一部に相補的な関係がある [3]．相補的な関係を活用する方法として，逆翻訳確率を用いた方法および単語対応を用いた方法のどちらも訳抜けしていると判別した場合に訳抜けしていると判別する．この方法を Combi と呼ぶ．

2.3　テストデータへのラベル付与方法

テストデータの各単語には訳出されていることを表すラベル t を付与する（図 2(b)）．ただし，訓練データ中で，ラベル t が付与されず，ラベル D しか付与されなかった単語については，テストデータでもラベル D を付与する．これは訓練時と推論時との不整合を削減するためである．

3　実験

日本語と英語の報道記事から自動対応付けして得られた対訳文対[*2]を用いて日英翻訳の実験を行った．

3.1　データ

訓練データは，時事通信社の日英記事（2011-2017）から自動抽出した対訳 23 万文対，および読売新聞の日本語記事と Daily Yomiuri の英語記事（2007-2017）から自動抽出した対訳 57 万文対である．文長が 100 単語以内の訓練データをモデルの学習に用いた．

テストデータと参照訳として，時事通信社の日英記事（2018）から自動抽出した対訳文から，主な内容が一致している日英対訳 1,912 文対を人手で選択し，日本語文をテストデータ，英語文を参照訳とした．開発データとして，時事通信社の日英記事（2018）でテストデータとは別の記事から，テストデータと参照訳の構築と同じ方法により 497 文対を抽出した．開発データは，訳抜け検出での閾値の選択に用いた．検証データとして，時事通信社の日英記事（2018）から対訳 1,000 文対を自動抽出した．検証データは NMT の訓練の早期停止に用いた．

3.2　設定

3.2.1　前処理

日本語文は Cabocha[*3]（IPA 辞書）で形態素解析し，単語分割と品詞付与を行った．連続した数字は 1 単語にまとめ，英語文は Lookahead タガー[*4]でトークン化と品詞付与を行った．さらに BPE[*5]により低頻度語を分割し

[*2] 本稿での文対は正確にはセグメント対で，目的言語側の 1 セグメントには複数文が含まれる場合がある．
[*3] https://taku910.github.io/mecab/
[*4] https://www.logos.ic.i.u-tokyo.ac.jp/ tsuruoka/lapos/
[*5] https://github.com/rsennrich/subword-nmt

た．BPE の設定には WAT2018 ベースラインシステムの設定を用いた．[*6]

3.2.2 ラベル推定

[9] の手法による単語対応の計算には，1 層の双方向 LSTM エンコーダ・デコーダ NMT を用いた．デコーダーの構成は [5] で，アテンション機構は MLP[1] である．primitiv フレームワーク[*7]を用いて実装した．[9] に従って，アテンション確率が GIZA++ のアラインメントに近くなるようにバイアスをかけた．原言語から目的言語への NMT モデルと目的言語から原言語への NMT モデルを訓練した．

比較のために GIZA++ と grow-diag-final-and ヒューリスティックによる単語対応もラベル推定に用いた．この方法を GIZA++ alignment と呼ぶ．

逆翻訳確率は，上記で訓練した目的言語から原言語への NMT モデルを用いて計算した．

閾値は，開発データで BLEU が最大になる値を選択した．[*8]

3.2.3 翻訳

NMT の手法には Transformer [12] を用い，実装には sockeye[*9]を用いた．また，比較手法として sockeye の RNN ベースの NMT も用いた．バッチサイズは 2,048 words とし，その他の設定はデフォルト設定を用いた．スコアに長さの正規化を適用し，ビームサイズ 5 の訳文候補から $(\sum_i \log p(y_i))/l_y$ でスコアを計算した．ここで l_y は出力長を表す．学習にはばらつきがあり，まれにパラメータの最適化に失敗して BLEU スコアが低い外れ値が存在する．一方，BLEU スコアが高くなる外れ値は基本的にないと考えられる．そのため，5 回学習して，最も高い BLEU スコアの結果を選択した．

3.3 結果

表 1 に BLEU の brevity ペナルティの計算で用いる Length Ratio（出力長／参照訳長），および BLEU-4 スコアを示す．RNN，Transformer 共に Length Ratio が 0.93 程度で参照訳より出力長が短いことが分かる．それに対して提案手法では，参照訳の長さに近い出力長となっていることが分かる．このことから，訳出の不足が減っていることが期待できる．また，BLEU スコアは，Proposed (Combi) で Transformer に比べて 0.79 BLEU ポイント向上した．

表 1　Length Ratio (L-ratio) と BLEU スコア

	L-ratio	BLEU (%)
RNN	0.926	21.26
Transformer	0.927	24.33
Proposed (GIZA++ alignment)	1.026	24.28
Proposed (NMT alignment)	1.021	25.03
Proposed (BT probability)	0.984	25.12
Proposed (Combi)	1.001	25.12

表 2　内容語の訳抜け率と繰り返し率の評価結果 (%)

	訳抜け	繰り返し
Transformer	12.9	0.35
Proposed (Combi)	9.7	0.06

表 3　ラベル D の推定の評価結果 (%)

	Precision	Recall
GIZA++ alignment	33.6	72.1
NMT alignment	44.0	41.6
BT probability	38.0	30.8
Combi	54.3	26.4

Transformer と，提案手法の中で BLEU スコアが最も高く Length Raito が 1 に近い Proposed (Combi) について，ランダムに選択した 200 文について内容語の訳抜け率と繰り返し率を人手で調べた．結果を表 2 に示す．Transformer では 12.9% の内容語が訳抜けしていたが，提案手法では訳抜けが 9.7% に減少したことが確認された．減少率は 24.8% $((12.9-9.7)/12.9)$ である[*10]．同じ内容を 2 回以上訳出する繰り返しはベースラインと提案手法のいずれの結果でも少なく，提案手法で増加するということはなかった．

提案した枠組みによる効果は，訓練データに対するラベルの推定性能に依存する．そこで，ラベルの推定性能を調べた．訓練データから，時事通信社の対訳 200 文対と読売新聞の対訳 300 文対をランダムに抽出し，訳抜け部分に人手でタグ付けした．抽出した訓練データでの内容語の訳抜け率は 12.2% であった．正解タグに対する自動付与したラベル D の Precision と Recall を表 3 に示す．Precision が最も高かった Combi で 54.3% であり，推定が難しいことが分かる．[*11]

提案手法において，訓練データに付与したラベルの誤りの影響について考察する．GIZA++ alignment を用いた結果において，多義語の訳語選択の性能低下が観察さ

[*6] http://lotus.kuee.kyoto-u.ac.jp/WAT/WAT2018/baseline/dataPreparationBPE.html
[*7] https://github.com/primitiv/primitiv
[*8] 単語対応の閾値は {0.02, 0.01, 0.005, 0.002, 0.001} から，逆翻訳確率の閾値は {0.1, 0.05, 0.02, 0.01, 0.005, 0.002, 0.001} から選択した．
[*9] https://github.com/awslabs/sockeye
[*10] データが異なる為に単純な比較はできないが，リランキングを行う Reconstruction を用いた論文 [10] で報告された減少率は 11.0% $((18.2-16.2)/18.2)$ である．
[*11] ラベル推定のモデルの訓練データにノイズが多いことが推定を難しくしている原因の 1 つと考えられる．

れた．多義語の原言語単語の訳が訳語 1 と訳語 2 の 2 つがあり，訓練データで訳語 1 の頻度が高く，訳語 2 の頻度が低い状況で，訳語 2 との単語対応が得られない場合があった．この場合，訳語 1（例：China）を含む対訳文ではその原言語単語（例：中国）にはラベル t が付与され，訳語 2（例：Chugoku［日本の地域］）を含む対訳文ではその原言語単語にはラベル D が付与されることになる．テストデータではラベル t を付与することで訳語 2 に翻訳すべき場合でも常に訳語 1 が出力されてしまった．このような訳語選択性能の低下が，GIZA++ alignment で BLEU スコアが向上しなかった原因と考えられる．このことから，訳抜け検出の Precision が重要であることが分かる．

一方で，Combi の Recall は 26.4% にとどまっており，推定誤りを増やさずに Recall を改善できれば，提案手法の枠組みは訳抜けの低減効果の向上が期待できる．

4 関連研究

訓練時と推論時とのギャップを埋める機械翻訳の研究として，ドメインアダプテーションの研究 [2]，および翻訳履歴の訓練時と翻訳時のギャップを扱った研究 [14] がある．これらの研究とは訓練時と推論時とのギャップを埋めるというアプローチに共通点があるが，課題が本稿の対象とは異なる．

訳抜けを低減させる研究も提案されている．[11, 7, 16, 6, 15] の手法は，訓練データの特徴を学習する手法で，訓練時と推論時とのギャップを埋める手法ではない．このため，訓練データに多くの訳抜けが含まれている場合，訳抜けを学習してしまうという問題を解決できない．[13, 10, 3, 4] も NMT でギャップを埋める手法ではなく，後処理でリランキングしている．それに対して，提案手法は，NMT で訓練時と推論時とのギャップを埋める手法である．リランキングにより訳抜けの少ない候補を選択する手法 [13, 10, 3] は，候補の中に訳抜けのない出力が含まれていないと訳抜けがない訳を出力することができないという課題がある．提案手法は，リランキングの手法ではなく，訳抜けが少ない翻訳候補を直接生成する．提案手法は，リランキングする翻訳候補の生成に使うこともできるため，これらの手法と競合するのではなく，リランキングする手法と組み合わせて利用できる．

5 おわりに

訳抜けを多く含むデータで学習し，訳抜けしていない翻訳を出力するという訓練時と推論時とのギャップを埋める手法を提案した．提案手法は，訓練データ中の訳抜けを検出して，訳出されていない部分と訳出されている部分をラベルを用いて区別して学習し，翻訳時は訳出されている部分から学習した知識を主に使って翻訳することで訳抜けを抑える．日英の報道記事を用いて日英翻訳の実験を行い，24.8% の訳抜け減少を確認した．提案手法の枠組みは，ラベルの推定性能を向上させられれば効果の向上が期待できる．

謝辞

本研究成果の一部は，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．

参考文献

[1] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In *Proceedings of ICLR*, 2015.
[2] C. Chu, R. Dabre, and S. Kurohashi. An empirical comparison of domain adaptation methods for neural machine translation. In *Proceedings of ACL*, pp. 385–391, 2017.
[3] I. Goto and H. Tanaka. Detecting untranslated content for neural machine translation. In *Proceedings of the First Workshop on NMT*, pp. 47–55, 2017.
[4] Y. Li, T. Xiao, Y. Li, Q. Wang, C. Xu, and J. Zhu. A simple and effective approach to coverage-aware neural machine translation. In *Proceedings of ACL*, pp. 292–297, 2018.
[5] T. Luong, H. Pham, and C. D. Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of EMNLP*, pp. 1412–1421, 2015.
[6] F. Meng, Z. Tu, Y. Cheng, H. Wu, J. Zhai, Y. Yang, and D. Wang. Neural machine translation with key-value memory-augmented attention. In *Proceedings of IJCAI*, pp. 2574–2580, 7 2018.
[7] H. Mi, B. Sankaran, Z. Wang, and A. Ittycheriah. Coverage embedding models for neural machine translation. In *Proceedings of EMNLP*, pp. 955–960, 2016.
[8] M. Morishita, J. Suzuki, and M. Nagata. NTT neural machine translation systems at WAT 2017. In *Proceedings of WAT*, pp. 89–94, 2017.
[9] J.-T. Peter, A. Nix, and H. Ney. Generating alignments using target foresight in attention-based neural machine translation. *PBML*, 108(1):27–36, 2017.
[10] Z. Tu, Y. Liu, L. Shang, X. Liu, and H. Li. Neural machine translation with reconstruction. In *Proceedings of AAAI*, 2017.
[11] Z. Tu, Z. Lu, Y. Liu, X. Liu, and H. Li. Modeling coverage for neural machine translation. In *Proceedings of ACL*, pp. 76–85, 2016.
[12] 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 *Proceedings of NeurIPS*, pp. 5998–6008. 2017.
[13] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint*, abs/1609.08144, 2016.
[14] W. Zhang, Y. Feng, F. Meng, D. You, and Q. Liu. Bridging the gap between training and inference for neural machine translation. In *Proceedings of ACL*, pp. 4334–4343, 2019.
[15] Z. Zheng, S. Huang, Z. Tu, X.-Y. Dai, and J. Chen. Dynamic past and future for neural machine translation. In *Proceedings of EMNLP-IJCNLP*, pp. 931–941, 2019.
[16] Z. Zheng, H. Zhou, S. Huang, L. Mou, X. Dai, J. Chen, and Z. Tu. Modeling past and future for neural machine translation. *TACL*, 6:145–157, 2018.