ニューラル機械翻訳における複数モデルの利用

Usage of Multiple Models in Neural Machine Translation

国立研究開発法人情報通信研究機構　先進的音声翻訳研究開発推進センター主任研究員
今村　賢治

2004年奈良先端科学技術大学院大学情報科学研究科情報処理学専攻博士課程修了。1985年日本電信電話株式会社。2014年株式会社ATR-Trekより NICTに出向。機械翻訳の研究に従事。

国立研究開発法人情報通信研究機構　先進的音声翻訳研究開発推進センター副センター長
隅田　英一郎

1999年京都大学大学院博士（工学）取得。1982年日本アイ･ビー･エム。2002年ATR。2010年NICT（2016年同フェロー）。機械翻訳の研究に従事。

1　はじめに

近年主流となっているニューラル機械翻訳（neural machine translation; NMT）[1][2]は、単一のモデルでも高い翻訳性能を示すことが多いが、複数のモデルを用いると、さらに良質な翻訳が可能となる。複数のモデルを使う代表的な方法としては、アンサンブル[3]とリランキング（たとえば[4]）が知られている。このうち、アンサンブルは、複数のモデルで独立にエンコード、デコードを行い、その出力を平均する（2.1節参照）。一方、リランキングは、あるモデルAで翻訳仮説をN個生成（Nベスト翻訳）し、それを別のモデルBで再スコアリング、最もスコアの高い仮説を出力する（2.2節参照）。両者には、表1に示すようなメリット、デメリットがあり、これらの特徴を考慮して組み合わせることにより、翻訳精度の向上が期待できる。

[5]はアンサンブルとリランキングを併用した複数モデル利用法を提案した。彼らは、文の尤度をそのまま用いて方式比較を行ったが、方式や設定によって、尤度最大の翻訳文の長さは異なる。そのため、文の長さを調整（正規化）してから、方式比較をすべきである。本稿では、中規模の医療コーパスを例にとり、複数モデル利用時の翻訳精度を、長さ正規化を行ったうえで測定した結果について報告する。

表1　アンサンブルとリランキングのメリット・デメリット

	メリット	デメリット
アンサンブル	・すべての翻訳仮説が候補となる ・並列処理による高速化が可能	・語彙またはデコード方向が異なるモデルは併用不可 ・全モデルをGPUに載せた方がよい
リランキング	・言語対が同じなら、どのようなモデルも併用可 ・Nベスト生成、再スコアリングそれぞれのモデルがGPUに載れば、高速に動作	・Nベストリストにない候補は選択できない ・処理は逐次的

2　方式

本稿で仮定するニューラル機械翻訳は、再帰ニューラルネットワーク（recurrent neural network; RNN）ベースのエンコーダー・デコーダー方式である。sequence-to-sequenceモデルとも呼ばれる。エンコーダーは、入力文を状態と呼ばれる固定長の実数ベクトルに符号化する。デコーダーは、状態から、翻訳文を1単語ずつ

生成してゆく。最終的にデコーダーが文末記号(end-of-sentence; EOS)を生成したら翻訳を終了する。

2.1 アンサンブル

デコーダーが翻訳文を 1 単語ずつ生成する過程で、全単語の生成確率（事後確率分布）を保持した単語出力分布が生成される。これは、語彙サイズの次元を持った実数ベクトルである。通常は、単語出力分布から、確率の高い単語が選択され、ビームサーチへと送られる。一方アンサンブルは、複数のモデルの単語出力分布を平均する方法である。ビームサーチは、この平均化ベクトルを元に行われる。なお、モデルの訓練時は、複数のモデルはそれぞれ独立に、通常どおり訓練する。

1 モデルにおける出力単語の選択を式(1)であらわすと、アンサンブルにおける出力単語の選択は、式(2)となる。

$$\hat{y}_t = \underset{y_t}{\operatorname{argmax}} \Pr(y_t | \boldsymbol{y}_{<t}, \boldsymbol{x}; M) \quad (1)$$

$$\hat{y}_t = \underset{y_t}{\operatorname{argmax}} \frac{1}{J} \sum_{j=1}^{J} \Pr(y_t | \boldsymbol{y}_{<t}, \boldsymbol{x}; M_j) \quad (2)$$

ただし、y_t は時刻 t における出力単語、$\boldsymbol{y}_{<t}$ は、先頭から t 直前までの出力単語の履歴、$\boldsymbol{x}$ は入力単語列、M はモデル（M_j は複数モデルの j 番目）、J はモデル数（アンサンブル数）である。

アンサンブルは、単語出力分布を平均するため、目的言語の語彙はすべてのモデルで同じものを使用する必要がある。また、ビームサーチはアンサンブルの後に適用されるため、デコード方向（文頭から文末か、文末から文頭か）は、全モデルで一致していなければならないという制約がある。

2.2 リランキング

リランキングは、2 ステップの翻訳方式である。まず、モデル A で N ベスト翻訳を行い、翻訳仮説を N 個生成する。次に、別のモデル B で各翻訳仮説を再スコアリングする。最終的にモデル A によるスコアとモデル B によるスコアから、最も高い仮説を選択して出力する(図 1)。本稿では、対数尤度の算術平均を使用する。

リランキングは、言語対が合っていれば、どのようなモデルも利用できる。また、全体では 2 モデルを使用するにも関わらず、1 ステップで使用するモデルは 1 つなので、使用メモリはアンサンブルに比べると少ない。

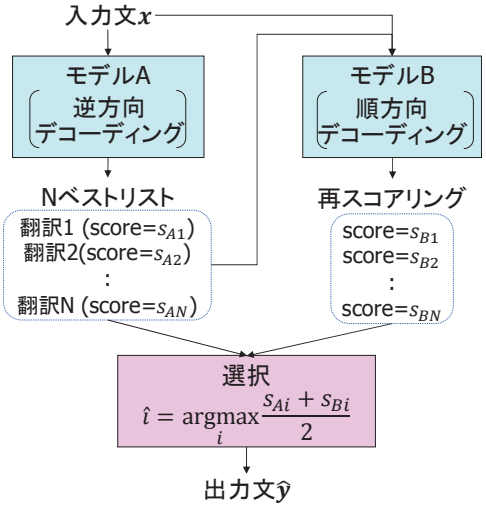

図 1　双方向リランキングの構成

しかし、良い仮説が N ベストリストに入らない場合は、翻訳品質は変わらないというデメリットもある。

2.3 両者の組み合わせ

本稿では、アンサンブルとリランキングのメリット、デメリットを考慮して、組み合わせるために、全体をリランキングで構成する。全体構成を図 1 に示す。N ベスト生成、再スコアリングそれぞれでアンサンブルを使用して、複数モデルを組み合わせる。リランキングは、1 つの処理におけるメモリ使用量がアンサンブルの半分であるため、この組み合わせでより多くのモデルを組み合わせることができる。

今回、すべてのモデルについて、語彙は同一のセットを使用する。また、各モデルは、ランダムシードを変えて学習した同一構成のものを使用する。

リランキングでは、N ベスト生成と再スコアリングで、デコード方向が異なるモデル（アンサンブルでは組み合わせ不可能）を使用することができる。これを双方向リランキングと呼ぶ。具体的には、N ベスト生成では、文末から文頭方向（逆方向）にデコーディングを行い、再スコアリングでは、文頭から文末（順方向）にデコーディングを行う。そして、双方向の対数尤度を平均し、最大の仮説を出力する。

3 実験

本節では、提案方法の効果を、日英翻訳の翻訳品質で評価する。

3.1　実験設定

コーパス：本稿で用いたコーパスは、内部開発の医療コーパスである。コーパスサイズを表２に示す。これは、病院等における患者とスタッフの会話を、ライターが作文したものである。日本語で作成した疑似対話文を英語に翻訳することで作成した。

表２　コーパスサイズ

種別	文数	サブワード数（タイプ数）
訓練	238,214	日：約 381 万語(20,332 タイプ)
開発	1,000	英：約 325 万語(21,049 タイプ)
テスト	1,000	

コーパスのすべての文について、Unicode の NFKC 正規化を行い、日本語は MeCab[6]、英語は Moses ツールキット[7] の tokenizer で単語分割した。英語に関しては、Moses の truecaser で単語正規化を行った。さらに、日本語、英語双方について、バイトペア符号化[8] を用いて、約２万種のサブワードに分割した。

翻訳システム：翻訳システムには、OpenNMT[9]（LUA 版）を用いた。エンコーダーは２層双方向 LSTM（500＋500 次元）、デコーダーは２層 LSTM（1,000 次元）、アテンションは[10]のグローバルアテンションを使用した。学習は確率的勾配降下法（stochastic gradient descent; SGD）、学習率 1.0 で 10 エポック、その後学習率を半減させながら６エポック学習した。

なお、2.3 節で述べた方式を実現するため、OpenNMT に以下の改造を施した。

- 翻訳器をアンサンブル化した。
- デコーダーの学習および翻訳を、文末から文頭方向に行えるようにした。

翻訳は、ビーム幅 10 で 10 ベスト翻訳を行い、双方向リランキングなどを行った。

評価：評価は、バイトペア符号化を解除した単語列について（detruecasing は行っていない）、BLEU[11] を用いて行った。

3.2　単語ペナルティと長さ比

ニューラル機械翻訳は、ビームサーチによって、文の尤度が最大の仮説を出力するが、これは、各単語の事後確率の総積であるため、単語数が少ない仮説の方が有利となる[12]。つまり、本来なら翻訳文は、参照訳と（ほぼ）同じ長さになるべきだが、短めの訳になる場合が多い。

この現象は、異なるモデルの翻訳品質を比較する場合に問題となる。つまり、BLEU のように翻訳文と参照訳との長さの差異もスコアに取り入れている評価指標の場合、モデルの品質ではなく、単なる翻訳文の長さの違いがスコアの差になってしまう場合がある。この問題を避けるためには、文の尤度を単語数で正規化するのが有効である。

本稿では、単語ペナルティ方式による正規化[13] を行う。この方式では、文の尤度を以下の式で表す。

$$ll_{bias}(\boldsymbol{y}|\boldsymbol{x}) = \sum_{t}^{J} \log \Pr(y_t|\boldsymbol{y}_{<t},\ \boldsymbol{x}) + WP \cdot T \qquad (3)$$

ただし、$ll_{bias}(\boldsymbol{y}|\boldsymbol{x})$ は長さ正規化済みの文の対数尤度、WP は単語ペナルティ（$WP \geq 0$）、T は出力単語数である。単語ペナルティ WP は正の値を設定するが、大きいほど長い翻訳文が出力されるようになる[1]。

図２は、順方向モデルを単独で使用した場合（アンサンブル数 1）の単語ペナルティと BLEU スコアおよび長さ比の関係を表したグラフである。長さ比とは、テスト／開発セットの翻訳文の単語数の総和を、参照訳の単語数の総和で割ったものである。ただし、バイトペア符号化を使用しているため、サブワード数で長さ比を算出した。

このグラフを見ると、BLEU スコアは、単語ペナルティを変化させると、1 ポイント程度変化する。そして、

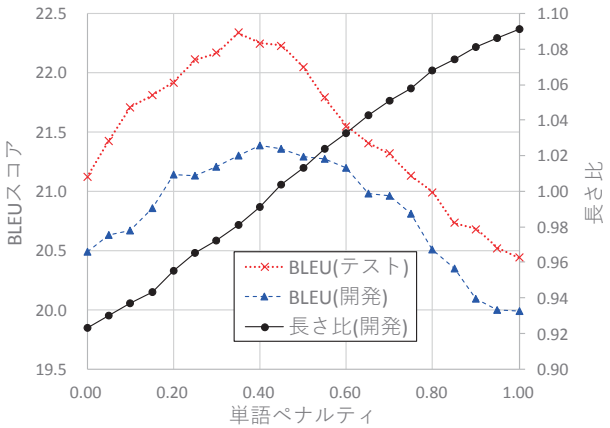

図２　単語ペナルティと BLEU スコア、および長さ比の関連

1　他にも、たとえば marian NMT[14] では、長さ正規化は以下の式で行っている。式(3)と同様に、WP を大きくした場合、より長い仮説が選ばれやすくなる。

$$ll_{norm}(\boldsymbol{y}|\boldsymbol{x}) = \frac{\sum_t \log \Pr(y_t|\boldsymbol{y}_{<t}, \boldsymbol{x})}{T^{WP}} \qquad (4)$$

長さ比が 1.0 付近のとき、最大となる。これは、長さ比が 1.0 の時は、BLEU の簡潔ペナルティがなく、かつ n グラム正解率も高く保持されるためである。

そこで本稿では、長さ比が一定になるように、単語ペナルティを調整し、精度を比較する。なお、サブワードから単語に戻す際にシンボル数が変化するため、長さ比 1.0 より少しずれる。本稿では、そのマージンを 0.02 程度とし、サブワードレベルでの長さ比が 0.98 になるように単語ペナルティを調整する。

3.3 複数モデルの効果

図 3 は、アンサンブル数を変えたときの BLEU スコアの変化を表す。なお、双方向リランキングは、N ベスト生成、再スコアリングでそれぞれアンサンブルモデルを使用するので、使用する総モデル数は、アンサンブル数の倍となっている。

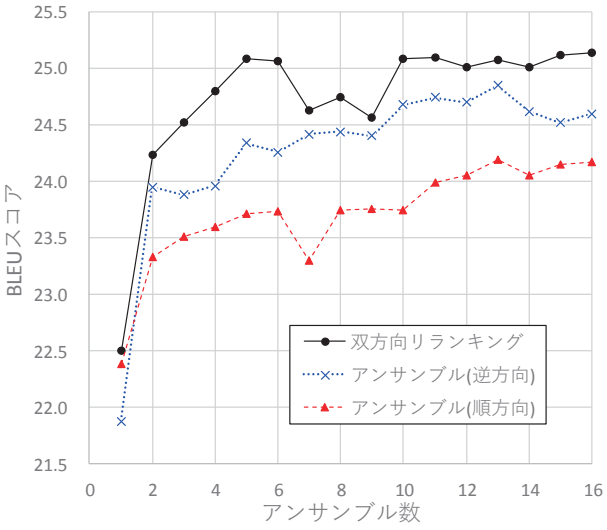

図 3　モデルの数と BLEU スコア

基本的には、アンサンブル、双方向リランキングともに、モデル数を増やすと BLEU スコアは向上する。ちなみに、BLEU スコアが最高となったモデル数は、順方向および逆方向アンサンブルでは 13 モデル、双方向リランキングでは 16 アンサンブル（のべ 32 モデル）であった。

このコーパスでは、逆方向モデルの方が、順方向モデルより BLEU スコアが高い傾向があるが、他のデータでも成り立つ傾向かは不明である。少なくとも、デコード方向を変えると、翻訳品質が変わることがあると言える。

双方向リランキングは、アンサンブル単独よりも BLEU スコアは高くなった。双方向リランキングは、アンサンブル単独の 2 倍のモデルを使用するため、同じモデル数同士で比較したが、それでもこのデータでは双方向リランキングが高かった。

4 まとめ

本稿では、アンサンブルとリランキングを併用した、複数モデルの利用法について述べた。アンサンブルもリランキングも、複数モデルを使うことで、BLEU スコアは向上する。本稿では、のべ 32 モデルを使用したときに BLEU スコアが最高となり、まだ飽和していないことが分かった。

多数のモデルを使用すれば、翻訳品質を向上させられることが分かったため、今後は単独モデルの精度向上に注力する。

謝辞

本研究は総務省の情報通信技術の研究開発「グローバルコミュニケーション計画の推進 - 多言語音声翻訳技術の研究開発及び社会実証- Ⅰ．多言語音声翻訳技術の研究開発」の一環として行われました。

参考文献

[1] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In Proceedings of Advances in Neural Information Processing Systems 27 (NIPS 2014), pages 3104-3112.

[2] Dzmitry Bahdanau, KyungHyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In Proceedings of International Conference on Learning Representations (ICLR 2015).

[3] Lars Kai Hansen and Peter Salamon. 1990. Neural network ensembles. IEEE Transactions on Pattern Analysis and Machine Intelligence, 12 (10): 993-1001.

[4] Franz Josef Och, Daniel Gildea, Sanjeev Khudanpur, Anoop Sarkar, Kenji Yamada,

Alex Fraser, Shankar Kumar, Libin Shen, David Smith, Katherine Eng, Viren Jain, Zhen Jin, and Dragomir Radev. 2004. A smorgasbord of features for statistical machine translation. In HLT-NAACL 2004: Main Proceedings, pages 161–168, Boston, Massachusetts, USA

[5] 今村賢治、隅田英一郎. 2017. 双方向リランキングとアンサンブルを併用したニューラル機械翻訳における複数モデルの利用法. 情報処理学会研究報告, 2017-NL-233, No.9, 10 月.

[6] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to Japanese morphological analysis. In Proceedings of EMNLP 2004, pages 230–237, Barcelona, Spain.

[7] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, pages 177–180, Prague, Czech Republic.

[8] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1715–1725, Berlin, Germany.

[9] Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In Proceedings of ACL 2017, System Demonstrations, pages 67–72, Vancouver, Canada.

[10] Thang Luong, Hieu Pham, and D. Christopher Manning. 2015. Effective approaches to attention-based neural machine translation. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1412–1421, Lisbon, Portugal.

[11] Kishore Papineni, Salim Roukos, ToddWard, andWei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), pages 311–318, Philadelphia, Pennsylvania, USA.

[12] Makoto Morishita, Jun Suzuki, and Masaaki Nagata. 2017. NTT neural machine translation systems at WAT 2017. In Proceedings of the 4th Workshop on Asian Translation (WAT2017), pages 89–94, Taipei, Taiwan.

[13] Yusuke Oda, Katsuhito Sudoh, Satoshi Nakamura, Masao Utiyama, and Eiichiro Sumita. 2017. A simple and strong baseline: NAIST-NICT neural machine translation system for WAT2017 English-Japanese translation task. In Proceedings of the 4th Workshop on Asian Translation (WAT2017), pages 135–139, Taipei, Taiwan.

[14] Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, Andr'e F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in c++. In Proceedings of ACL 2018, System Demonstrations, pages 116–121.