Vol.2017-NL-233 No.9
2017/10/24

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

今村 賢治[1,a)]　隅田 英一郎[1]

概要：本稿では，ニューラル機械翻訳における複数モデル利用法について提案する．提案方式は，アンサンブルによって複数モデルを使用し，さらにデコード方向が異なるモデルをリランキング法（双方向リランキングと呼ぶ）によって組み合わせる．
小規模データセットによる実験では，使用モデル数を増加させると翻訳品質は向上し，のべ 32 モデルまで増加させても，翻訳品質は悪化しなかった．また，データセットによってはさらなる向上の余地があった．大規模データセットの実験では，6 アンサンブルモデルを双方向リランキングによって組み合わせることで，単一モデルに比べ，BLEU スコアが 1.59〜3.32 ポイント向上した．

Usage of Multiple Models by Ensemble and Bi-directional Reranking in Neural Machine Translation

Kenji Imamura[1,a)]　Eiichiro Sumita[1]

1. はじめに

近年，機械翻訳はエンコーダー・デコーダー方式（Sequence-to-Sequence 学習）に基づくニューラル機械翻訳 (NMT) [1], [2] が主流となってきている．この翻訳方式は，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になる．

複数のモデルを使う代表的な方法として，アンサンブル [3] とリランキング（たとえば [4]）が知られてる．アンサンブルは，入力文を複数のモデルでそれぞれエンコード，デコードし，その出力である単語出力分布を平均する（2.2 節参照）．一方，リランキングは，モデル A で生成した N ベストの翻訳結果を，別のモデル B で再スコアリングし，最も高いスコアの翻訳を採用する（2.3 節参照）．

本稿では，アンサンブルとリランキングを併用した，ニューラル機械翻訳における複数モデルの利用法を提案する．アンサンブルとリランキングにはそれぞれ，**表 1** に示すようなメリットとデメリットがある．これらの特徴を考慮して，複数のモデルを最大限使用するのが，本稿の目的である．

本稿では，まず，小規模なデータにおいて以下の特性を検証し，その後，大規模データに適用する．

- いくつの組み合わせまでが翻訳品質向上に寄与するか．
- モデル数が同じ場合，どの翻訳品質がよいか．ただし，今回は翻訳時間は考慮せず，翻訳品質だけを評価する．

以下，第 2 節では，アンサンブルとリランキングの詳細を説明し，第 3 節で，本稿の提案方式である組み合わせ方法を説明する．第 4 節では，小規模データセット（時事コーパスと，内部開発の医療コーパス）を用いて，アンサンブルとリランキングの特性を評価する．第 5 節では，ASPEC (Asian Scientific Excerpt Corpus)[5] の全データを用いて評価し，第 6 節でまとめる．

なお今回は，日英・英日翻訳，日中・中日翻訳しか行わなかったため，すべての言語対で一般性があるか，さらなる検証が必要であるが，一つのケーススタディとしては有用であると考える．

[1] 国立研究開発法人 情報通信研究機構
National Institute of Information and Communications Technology
[a)] kenji.imamura@nict.go.jp

	利点	欠点
アンサンブル	• すべての翻訳仮説が候補となる	• 出力層が異なるモデルは併用できない（語彙，デコード方向に関して）
	• 並列処理による高速化が可能	• 全モデルを GPU に載せた方がよい
リランキング	• 言語対さえ合っていれば，どのようなモデルも併用できる • 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する	• N ベストリストに入らなかった候補は選択できない • 処理は逐次的

表 1 アンサンブルとリランキングの利点と欠点
Table 1 Pros and Cons of Ensemble and Reranking

2. アンサンブルとリランキング

2.1 ニューラル機械翻訳におけるデコーディング

現在のニューラル機械翻訳で主流となっているエンコーダー・デコーダー方式 [1], [2] は，入力文をエンコーダーによって，分散表現に符号化し，デコーダーによって文を生成する方式である．デコーダーは，アテンションと呼ばれる機構によって原文を参照し，原文と翻訳文の等価性を高めるようにしている．エンコーダー，デコーダーともに，LSTM (long short-term memory) や GRU (gated recurrent unit) ベースの再帰型ニューラルネットワーク (recurrent neural network; RNN) で構成されている．

デコーダーからは，出力文の各単語ごとに，単語の生成確率を保持した単語出力分布が出力される．これは，語彙サイズの次元を持った実数ベクトルで，各次元が単語を意味している．この単語出力分布から，最大確率を持つ出力単語が選択され，次の単語を予測する際のデコーダーへの入力となる．

上記は，貪欲法による 1 ベスト翻訳の場合の動作であるが，通常，翻訳器は複数の候補（出力単語の履歴とデコーダーの状態を含む）を保持し，ビームサーチによって最良な候補を選択している．ビーム幅を N とした場合，ビームサーチでは，まず，デコーダーが N 個の状態から，それぞれ単語出力分布を生成する．そして，各出力分布の上位 N 単語から（つまり，$N \times N$ 候補から）N 個の新しい候補を残す．最終的に，単語出力の対数確率の総和を翻訳文の尤度とし，尤度最大の候補を出力する．**図 1** は，デコーダーとビームサーチの関係を表す模式図である．

2.2 アンサンブル

アンサンブルは，ニューラルネットワークにおいて，同一のデータセットで訓練された複数のモデルを用い，その出力を平均する方法である [3]．ニューラル機械翻訳の場合，入力文を複数のモデルでそれぞれエンコード，デコードし，デコーダーが出力した単語出力分布を平均する．そして，この平均化ベクトルを元にビームサーチを行う．訓練時は，各モデルは単一モデルと同様に，通常どおり学習

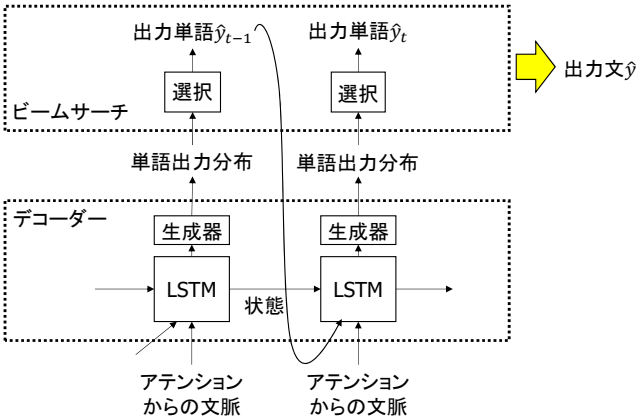


図 1 デコーダーとビームサーチの関係
Fig. 1 Relationship between Decoder and Beam Search

する．

単一モデルでの出力単語の選択を式 (1) で表すとすると，アンサンブルでの出力単語の選択は，式 (2) となる．なお，今回の平均化は幾何平均を使用する．

$$\hat{y}_t = \mathrm{argmax} \log Pr(y_t | y_1^{t-1}, \mathbf{x}; M) \quad (1)$$

$$\hat{y}_t = \mathrm{argmax} \frac{1}{J} \sum_{j=1}^{J} \log Pr(y_t | y_1^{t-1}, \mathbf{x}; M_j) \quad (2)$$

ここで，y_t は t 番目の出力単語，y_1^{t-1} は最初から $(t-1)$ 番目の出力単語の履歴，$\mathbf{x}$ は入力単語列全体，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．

アンサンブル法は，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，ビームサーチは出力のアンサンブル後に適用されるため，アンサンブルで用いられるモデルのデコード方向（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．

2.3 リランキング

リランキング [4] は，2 ステップの翻訳方式である．最初に，あるモデル A で入力文を N ベスト翻訳する．続いて，別のモデル B で N ベストリストの翻訳文を評価し，再スコアリングする．その結果，最もスコアの高い翻訳文を採

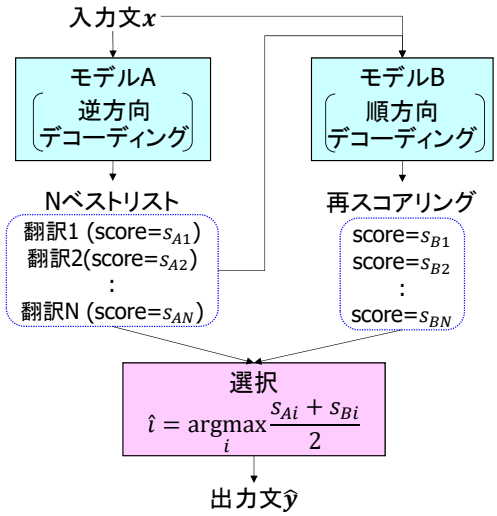


図 2 双方向リランキングの構成
Fig. 2 Structure of Bi-directional Reranking

用・出力する（**図 2**）．訓練時は，モデル A，モデル B ともに，独立に通常どおり学習する．再スコアリングの基準をどのように設定するかによって，翻訳文は変化する．本稿では，モデル A とモデル B の対数尤度の平均（算術平均）を使用する．

リランキングは，言語さえ合っていれば，どのようなモデルも利用できる点がメリットである．また，全体では 2 モデル使用するにもかかわらず，1 ステップで使用するモデルは 1 つなので，アンサンブルに比べると，使用メモリは半分で済むメリットがある．しかし，N ベストリストに良い翻訳が入らなかった場合は，翻訳品質が変わらないという欠点もある．

3. 提案方式: 双方向リランキングとアンサンブルの組み合わせ

アンサンブルとリランキングは，表 1 に示したようなメリット，デメリットが存在する．できる限り，方式のメリットを活かして両者を組み合わせるために，本稿では，全体をリランキングで構成する（図 2）．そして，N ベストリスト生成，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済むため，この構成でより多くのモデルを組み合わせることができる．

今回，アンサンブルを使用するため，すべてのモデルにおいて，語彙は固定とする．また今回は，個々のモデルは，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．

リランキングにおける N ベスト生成と再スコアリング用モデルでは，異なるデコード方向のモデルを使用し，アンサンブル不可能なモデルを組み合わせる．これを本稿では，双方向リランキングと呼ぶ．具体的には，文末から文頭方向（逆方向と呼ぶ）にデコーディングを行って N ベストリスト生成し，文頭から文末方向（順方向）デコーダーを使って再スコアリングを行う．そして，双方向の翻訳尤度を平均し，最大の翻訳仮説を出力する．

双方向リランキングは，[6] が提案した双方向デコーディング方式を，リランキングで実現したものである．目的言語の単語列を反転させて訓練，翻訳を行うだけで実現できるので，訓練プログラム，翻訳プログラムを変更する必要がほとんどない．

4. 小規模データによる実験

まず，アンサンブルと双方向リランキングの特性を把握するため，比較的小規模（約 20 万文）のデータで日英翻訳の実験を行う．

4.1 実験設定

4.1.1 コーパス:

表 2 は，今回使用したコーパスの一覧である．小規模コーパスとして，今回 2 種類のコーパスを利用した．一つめは，時事コーパスである．これは，ニュース記事のコーパスで，日本語記事と英語記事を文単位に自動対応付けすることで作成されている．日本語記事と英語記事は，必ずしも文単位に翻訳されたものではないため，対訳文も単語単位には対応づかない場合が多い．

二つめは，内部開発した医療コーパスである．これは，病院における患者とスタッフの会話を，ライターが作文した疑似対話コーパスである．日本語で作成した対話文を英語に翻訳することで作成した．

これら 2 つのコーパスの訓練セットそれぞれから，バイトペア符号化 [7] でサブワード分割ルールを獲得し，訓練，開発，テストセットに適用した．サブワードの種類数は，時事コーパスでは 3 万 4，5 千程度，医療対話コーパスでは，2 万強である．そして，サブワード数が 80 単語以下の文を使って訓練した．

4.1.2 前処理，後処理

表 3 は，今回の翻訳システムの概要である．

前処理については，コーパスのすべての文について，まず Unicode の NFKC 正規化を行った．次に，日本語は MeCab[8]，英語は Moses ツールキット [9] の Tokenizer，中国語は Stanford Word Segmenter(Chinense Penn Treebank モデル)[10] で単語分割した．そして，バイトペア符号化を用いてサブワード分割した．

後処理については，日本語及び中国語の出力文中の単語区切り記号を削除した．英語に関しては，Moses の DeTrueCaser，DeTokenizer を用いて，翻訳文を作成した．

コーパス	言語対	文数		サブワード数		備考
時事コーパス	日英	訓練:	199,905	日:	35,009	ニュース記事
		開発:	2,000	英:	33,934	サブワード数が 80 以下の文
		開発テスト:	2,000			
		テスト:	2,000			
医療コーパス	日英	訓練:	238,214	日:	20,327	病院等における疑似対話
		開発:	1,000	英:	21,043	サブワード数が 80 以下の文
		テスト:	1,000			
ASPEC	日英・英日	訓練:	2,977,320	日:	49,656	科学技術文献
		開発:	1,790	英:	49,776	サブワード数が 80 以下の文
		テスト:	1,812			
	日中・中日	訓練:	656,635	日:	49,654	科学技術文献
		開発:	2,090	中:	49,385	サブワード数が 80 以下の文
		テスト:	2,107			

表 2 コーパスの統計量
Table 2 Corpus Statistics

		日本語	英語	中国語
前処理	文字正規化	Unicode の NFKC 正規化		
	単語分割	MeCab	Moses ツールキット	Stanford Segmenter (CTB モデル)
	TrueCaser	–	Moses ツールキット	–
	バイトペア符号化	内部開発		
学習と翻訳	システム	OpenNMT (アンサンブルおよび逆方向デコーディング改造版)		
	エンコーダー	単語分散表現: 500 次元，2 層 Bi-LSTM (500 + 500 次元)		
	デコーダー	単語分散表現: 500 次元，2 層 LSTM (1,000 次元)		
	アテンション	グローバルアテンション		
	学習	ミニバッチサイズ:64, SGD 最適化 (10+6 epochs), ドロップアウト:0.3		
	翻訳	ビーム幅:5 (4.2 節参照)		
後処理	DeTrueCaser	–	Moses ツールキット	–
	DeTokenizer	空白削除	Moses ツールキット	空白削除

表 3 実験システムの概要
Table 3 Summary of Experimental System

4.1.3 翻訳システム

翻訳システムは，OpenNMT[11][*1]をベースにした．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（1000 次元），アテンションはグローバルアテンション [12] を使用した．学習は，確率的勾配降下法 (stochastic gradient descent; SGD) で最適化した．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．

今回，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（逆方向）に向けて行えるようにした．

リランキングのための N ベストは，4.2 節の実験に基づき設定する．

4.1.4 評価

評価は，大文字小文字を区別した BLEU[13] で行った．評価のための単語分割器に関しては，日本語は JUMAN[14]，英語は Moses ツールキットの Tokenizer，中国語は Stanford Word Segmenter（Chinese Penn Treebank モデル）を使用した．これは，国際ワークショップ WAT (Workshop on Asian Translation) [15] の設定の一つと同じである．

[*1] http://opennmt.net/

4.2 最適な N ベストサイズ

ビームサーチで N ベストを出力するためには，ビーム幅は N 以上確保した方がよいが，この実験では，N ベストサイズとビーム幅は同じ値を使用した．

図 3 は，時事コーパスの開発テストセットにおける，N ベストサイズを変えたときの BLEU スコアである．順方向，逆方向デコーディングそれぞれと，両者を双方向リランキングしたときのスコアを示す．なお，この実験では，アンサンブルは使用せず，単一モデルだけを使用した．

順方向，逆方向デコーディング，双方向リランキングともに，N ベストサイズを変化させると，BLEU スコアも変化する．単一モデルである順方向，逆方向デコーディング

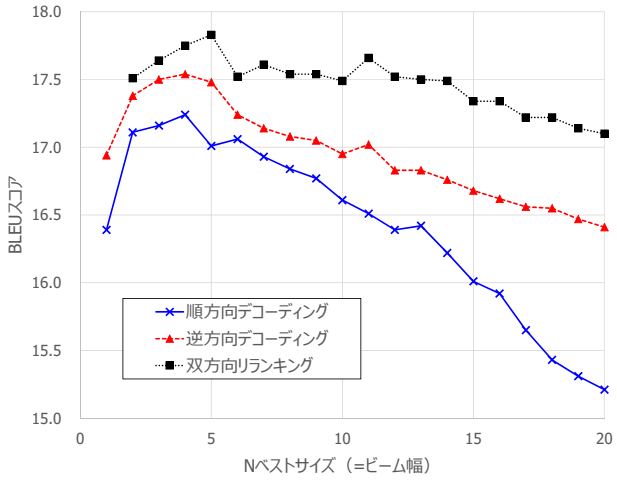


図 3 N ベストサイズごとの BLEU スコア
Fig. 3 BLEU Scores According to N-best Size

は，N ベストサイズが 4 の時に最高値を示し，N ベストサイズを大きくすると，BLEU スコアは悪化した．双方向リランキングは，N ベストサイズ 5 の時に最高値を示し，その後ゆるやかに低下した．

一般的にリランキングでは，良い翻訳仮説を含めるために，大きな N ベストサイズが望まれる．しかしこの実験では，比較的小さい N ベストサイズに単一モデルのピークが来ており，リランキングでも，小さな N ベストサイズで最高値となるという結果になった．これは，N ベストサイズ拡大によるカバレッジ向上よりも，単一モデルの精度低下の悪影響が大きくなったためと考えられる．以上の観察から，この後の実験はすべて N ベストサイズ 5 を使用する．

4.3 複数モデルの効果

時事コーパス，および医療コーパスにおけるアンサンブル単独（順方向デコーディングと逆方向デコーディング），および双方向リランキングの結果を，それぞれ**図 4**，**図 5** に示す．なお，リランキングのモデル総数は，実質アンサンブル単独の 2 倍になるため，アンサンブル数をベースにしたグラフを (a) に，モデル総数をベースにしたグラフを (b) に示す[*2]．なお，モデル数，アンサンブル数ともに，少ない設定から，一つづつ追加する方法（つまり漸増）で増加させた．したがって，モデル数が多い設定は，必ず少ない設定を包含している．

まず，アンサンブル数（モデル総数）と翻訳品質の関連を見ると，アンサンブル単独，双方向リランキングともに，基本的には，アンサンブル数を増加させると，BLEU スコアは向上する．ただし，増加率はだんだん低くなる傾向がある．時事コーパスの場合，16 アンサンブルでもまだ若干，BLEU スコアは向上している．医療対話コーパスでは，アンサンブル単独は，2〜6 モデルで収束しているが，双方向リランキングは 16 アンサンブルでもまだ向上している．

[*2] アンサンブル単独の場合は，モデル総数とアンサンブル数は同じ．

文献 [16] は，全モデルをアンサンブルするより，有効なモデルを取捨選択した方がよいと述べている．しかし本実験では，全モデルを使用しても，精度低下にはならなかった．結果，アンサンブル（順方向）の単一モデルにくらべ，16 アンサンブルの双方向リランキング（のべ 32 モデル使用）では，時事コーパスの BLEU スコアは 1.67 ポイント，医療コーパスのスコアは 2.58 ポイント向上した．

次に，アンサンブル法の順方向と逆方向を比較すると，医療コーパスの場合，逆方向デコーディングの方が順方向デコーディングより，多くの場合で高い BLEU スコアを示している．一方，時事コーパスでは，デコード方向によらず，翻訳品質はほぼ同じである．この結果は，コーパスや言語対などによって異なると思われるが，少なくともデコード方向によって翻訳品質は変わることがあることを示している．

最後に，グラフ (a) を見ると，双方向リランキングはほぼ常に，アンサンブル単体の精度を上回っている．この結果は，双方向リランキングはアンサンブルによる品質向上とは別の観点で翻訳品質を向上させられることを示している．今回は，デコード方向を変えたものを組み合わせているため，双方向デコーディング [6] と同様な効果が表れたものと思われる．

グラフ (b) は，リランキングのモデル総数が，アンサンブルの 2 倍であることを考慮したグラフであるが，これを見ても，双方向リランキングはアンサンブル法をほぼ常に上回っている．この実験では，モデル数が同じならば，双方向リランキングはアンサンブルより高い効果があった．

5. 大規模データを用いた実験

本節では，ASPEC の全データを用いて，日英，英日，日中，中日の翻訳実験を行う．

5.1 実験設定

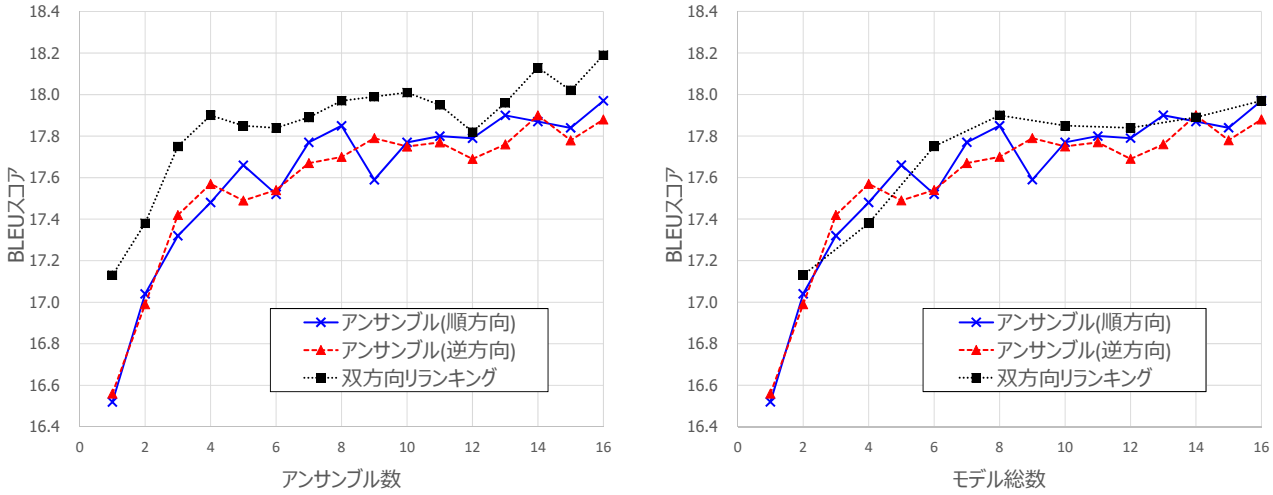
コーパスは，表 2 の ASPEC データセットを使用した．この訓練セットからバイトペア符号化によって約 5 万のサブワードに分割し，サブワード数が 80 単語以下の文を学習した．

使用システム，前処理，後処理など，その他の設定は，4 節のものと同じで，設定概要は，表 3 に示したとおりである．

5.2 実験結果

日英，英日翻訳の結果を**表 4** に，日中，中日翻訳の結果を**表 5** に示す．

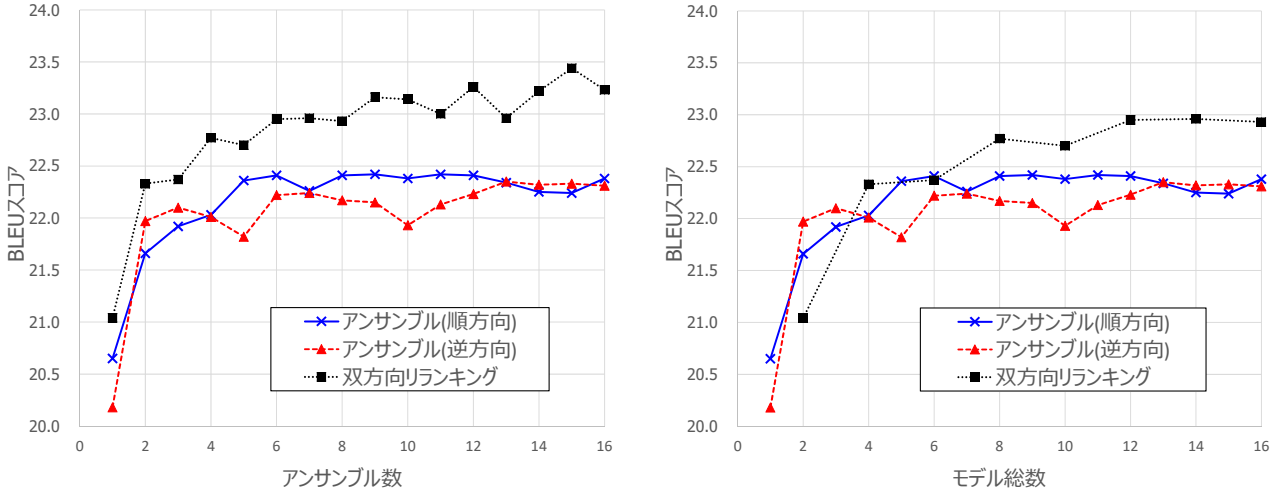
今回はリソースの関係で，6 アンサンブルまでしか試していないが，小規模データの場合と同様な傾向が得られた．つまり，アンサンブル単独，双方向リランキングともに，モデル数を増加させると，BLEU スコアが向上している．



(a) アンサンブル数と BLEU スコア　(b) モデル総数と BLEU スコア

図 4　時事コーパスによる複数モデル組み合わせ実験結果

Fig. 4　Results of Multiple Model Combination on the JIJI Corpus



(a) アンサンブル数と BLEU スコア　(b) モデル総数と BLEU スコア

図 5　医療対話コーパスによる複数モデル組み合わせ実験結果

Fig. 5　Results of Multiple Model Combination on the MED Corpus

そして，英日翻訳を除き，すべての言語対で，6 モデルアンサンブルを用いて，双方向リランキングした場合に最も高い BLEU スコアを示した．

結果，順方向の単一モデルに対する 6 モデルアンサンブルの双方向リランキングの BLEU スコアは，日英，英日，日中，中日翻訳それぞれで，+1.97，+3.32，+1.59，+2.58 ポイント向上した．

6. おわりに

本稿では，ニューラル機械翻訳における複数モデル利用法について提案した．提案方式は，アンサンブルによって複数モデルを使用し，さらに双方向リランキングによってデコード方向が異なるモデルを組み合わせる．

小規模データセットの実験では，16 アンサンブルまでモデルを増加させたが，基本的にはモデル数を増加させると翻訳品質は向上し，データによってはさらなる向上の余地があるという結果になった．

また，双方向リランキングでは，デコード方向が異なるモデルを用いてリランキングを行い，アンサンブル以上の翻訳品質向上を達成した．アンサンブルとリランキングを併用することで，さらに翻訳品質の向上が可能であり，ASPEC データを用いた実験では，単独モデルに比べ，BLEU スコアが 1.59〜3.32 ポイント向上した．また，デコード方向を変えると，翻訳品質が変化する場合があることがわかった．

アンサンブルもリランキングも，単一のモデルの性能を向上させれば，さらに全体性能を上げられる．今後は，単一モデルの性能向上に取り組むが，その際，単一モデルと複数モデルは，別に評価すべきと考えている．

現在のハードウェア環境では，（数，メモリ容量等）利用できる GPU に制約がある場合が多く，アンサンブルは実用的観点からは現実的ではないかもしれない．しかし，

	日英			英日		
アンサンブル数	アンサンブル (順方向)	アンサンブル (逆方向)	双方向 リランキング	アンサンブル (順方向)	アンサンブル (逆方向)	双方向 リランキング
1	24.79	24.72	25.34	36.85	38.20	39.10
2	25.60	25.40	25.89	38.37	38.69	39.41
3	26.17	25.62	26.08	38.95	39.23	39.87
4	25.89	25.77	26.26	38.97	39.37	40.03
5	25.94	26.06	26.37	39.19	39.55	**40.23**
6	26.21	26.29	**26.76**	39.13	39.26	40.17

表 4 日英・英日翻訳の BLEU スコア (ASPEC データ)

Table 4 BLEU Scores of Japanese-English and English-Japanese Translations (ASPEC Data)

	日中			中日		
アンサンブル数	アンサンブル (順方向)	アンサンブル (逆方向)	双方向 リランキング	アンサンブル (順方向)	アンサンブル (逆方向)	双方向 リランキング
1	33.64	33.60	34.10	44.26	44.13	45.10
2	34.67	34.22	34.77	45.59	45.52	46.20
3	34.75	34.64	34.98	45.88	45.93	46.53
4	34.75	34.64	34.98	46.13	46.10	46.55
5	35.02	34.81	35.18	46.27	46.36	46.69
6	35.27	34.95	**35.23**	46.55	46.31	**46.84**

表 5 日中・中日翻訳の BLEU スコア (ASPEC データ)

Table 5 BLEU Scores of Japanese-Chinese and Chinese-Japanese Translations (ASPEC Data)

ハードウェアの進展に伴って，制約がなくなる可能性もあると考えている．

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] Sutskever, I., Vinyals, O. and Le, Q. V.: Sequence to Sequence Learning with Neural Networks, *Proceedings of Advances in Neural Information Processing Systems 27 (NIPS 2014)*, pp. 3104–3112 (2014).

[2] Bahdanau, D., Cho, K. and Bengio, Y.: Neural Machine Translation by Jointly Learning to Align and Translate, *Proceedings of International Conference on Learning Representations (ICLR 2015)* (2015).

[3] Hansen, L. K. and Salamon, P.: Neural Network Ensembles, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 12, No. 10, pp. 993–1001 (1990).

[4] Och, F. J., Gildea, D., Khudanpur, S., Sarkar, A., Yamada, K., Fraser, A., Kumar, S., Shen, L., Smith, D., Eng, K., Jain, V., Jin, Z. and Radev, D.: A Smorgasbord of Features for Statistical Machine Translation, *HLT-NAACL 2004: Main Proceedings*, Boston, Massachusetts, USA, pp. 161–168 (2004).

[5] Nakazawa, T., Yaguchi, M., Uchimoto, K., Utiyama, M., Sumita, E., Kurohashi, S. and Isahara, H.: ASEPC: Asian Scientific Paper Excerpt Corpus, *Proceedings of the Tenth Edition of the Language Resources and Evaluation Conference (LREC-2016)*, Portoroz, Slovenia (2016).

[6] Liu, L., Utiyama, M., Finch, A. and Sumita, E.: Agreement on Target-bidirectional Neural Machine Translation, *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, California, pp. 411–416 (2016).

[7] 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 (Volume 1: Long Papers)*, Berlin, Germany, pp. 1715–1725 (2016).

[8] Kudo, T., Yamamoto, K. and Matsumoto, Y.: Applying Conditional Random Fields to Japanese Morphological Analysis, *Proceedings of EMNLP 2004*, Barcelona, Spain, pp. 230–237 (2004).

[9] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A. and Herbst, E.: Moses: Open Source Toolkit for Statistical Machine Translation, *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, Prague, Czech Republic, pp. 177–180 (2007).

[10] Chang, P.-C., Galley, M. and Manning, C. D. .: Optimizing Chinese Word Segmentation for Machine Translation Performance, *Proceedings of the Third Workshop on Statistical Machine Translation*, Columbus, Ohio, pp. 224–232 (2008).

[11] Klein, G., Kim, Y., Deng, Y., Senellart, J. and Rush, A. M.: OpenNMT: Open-Source Toolkit for Neural Machine Translation, *Proceedings of ACL 2017, System*

Demonstrations, Vancouver, Canada, pp. 67–72 (2017).

[12] Luong, T., Pham, H. and Manning, D. C.: Effective Approaches to Attention-based Neural Machine Translation, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, pp. 1412–1421 (2015).

[13] Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J.: Bleu: a Method for Automatic Evaluation of Machine Translation, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, Pennsylvania, USA, pp. 311–318 (2002).

[14] Kurohashi, S., Nakamura, T., Matsumoto, Y. and Nagao, M.: Improvements of Japanese morphological analyzer JUMAN, *Proceedings of the International Workshop on Sharable Natural Language*, pp. 22–28 (1994).

[15] Nakazawa, T., Mino, H., Ding, C., Goto, I., Neubig, G., Kurohashi, S. and Sumita, E.: Overview of the 3rd Workshop on Asian Translation, *Proceedings of the 3rd Workshop on Asian Translation (WAT2016)*, Osaka, Japan (2016).

[16] Zhou, Z.-H., Wu, J. and Tang, W.: Ensembling neural networks: Many could be better than all, *Artificial Intelligence*, Vol. 137, No. 1-2, pp. 239–263 (2002).