統計的機械翻訳とニューラル機械翻訳の混合 n ベストリランキング

Benjamin Marie　藤田 篤
情報通信研究機構

1 はじめに

ニューラル機械翻訳 (NMT) は，多くの翻訳タスクにおいて，句に基づく統計的機械翻訳 (PBSMT) を超える性能を示している．NMT による翻訳は，PBSMT による翻訳よりも流暢である場合が多いが，原文と意味が異なっていたり，訳し漏れや不要な語の繰り返しを生じたりする場合もある．また，対訳データが十分にない言語対やドメインなど，現実的な状況下で，PBSMT よりも性能が低いことが指摘されている [2, 17, 7] ．

より高品質な翻訳を生成するため，PBSMT と NMT を組み合わせる，次の 4 種類の手法が検討されてきた．

(a) 与えられた翻訳候補集合から，単語アラインメントに基づいて Confusion Network などを構成し，そこから改めて訳文をデコードする手法 [1, 4].
(b) PBSMT による n ベストの翻訳候補を NMT の尤度に基づいてリランキング する手法 [8].
(c) NMT による翻訳候補を PBSMT のモデルを用いてリランキングする手法 [19] (Phrase-based forced decoding; PBFD).
(d) PBSMT と NMT のパイプライン (事前翻訳) [11, 3].

我々は，PBSMT と NMT の各々を用いて生成した 2 つの翻訳候補集合を合わせてリランキングする．本稿では，我々が検討した種々の素性，および 4 つの翻訳タスクを用いた評価実験について述べる．

2 リランキング手法

n ベストの翻訳候補のリランキングは，探索空間における最適な部分空間のみを対象としつつ，デコード時には容易に参照できない素性も参照できるため，PBSMT の黎明期から用いられてきた [12]．このアプローチでは，翻訳候補の多様性が鍵となる [5]．上述の手法 (b), (c) は，単一のシステムから得られる翻訳候補のみを対象とするため，多様性に限界がある．上述の手法 (a) は，一般に複数の異なるシステムの翻訳候補を組み合わせることを想定している．元の翻訳候補集合には含まれない翻訳候補も探索空間に含めることになるため，性能が劣化する場合がある．

我々は，PBSMT と NMT の各々を用いて生成した n ベストの翻訳候補集合を合わせてリランキングする．単純なアプローチであるが，2 つの集合は性質が異なるため，解決が容易とは限らない．例えば NMT に基づく素性は，PBSMT から得た翻訳候補に対して低いスコアになりやすい．そこで，翻訳候補のリランキングに一般的に用いられている素性，特に PBSMT や NMT やデコード時には容易に参照できない素性を，PBSMT および NMT に基づく素性と組み合わせて用いる．

2.1 NMT に基づく素性

NMT の翻訳モデルを用いると，任意の原文に対する任意の翻訳の尤度を計算できる．一般に，異なる NMT のモデル，異なるエポックのモデル，異なる学習戦略・設定で得たモデルなど，複数のモデルをアンサンブルすることによって，より適切な尤度が得られる．

文頭から文末に向かって訳文を生成するモデル (left-to-right; L2R) で生成した翻訳候補をリランキングする際は，文末から文頭に向かって訳文を生成するモデル (right-to-left; R2L) による尤度が有用である [15].

2.2 PBSMT に基づく素性

NMT による翻訳候補に対して PBFD を適用し，そのスコアを素性として用いる [19]．PBSMT による翻訳候補については，デコード時の尤度をそのまま用いる．

PBFD によってフレーズの対応関係が求まれば，フレーズテーブルにおける翻訳確率の他，PBSMT で一般的に用いられる，歪みスコア，語彙化並び替えモデルのスコア，フレーズペナルティなどの素性も計算できる．さらに，それらと言語モデルのスコア，単語ペナルティを組み合わせて PBSMT の尤度も計算できる．

2.3 文レベルの翻訳確率

PBFD ではフレーズ単位の翻訳確率のみを用いるが，PBSMT や他の多言語処理タスクにおいて，語彙翻訳確率の有用性も報告されている．NMT は単語アラインメントを陽に特定しないので，次式の通り，原文 f と翻訳候補 e に含まれる全てのトークンの対についての翻訳確率を平均して文レベルの翻訳確率を得る．

$$P_{avg}(e|f) = \frac{1}{I}\sum_{i=1}^{I}\log\Big(\frac{1}{J}\sum_{j=1}^{J}p(e_i|f_j)\Big) \qquad (1)$$

ここで，I と J は各々，e と f におけるトークンの数を表す．$p(e_i|f_j)$ は f の j 番目の語 f_j に対する e の i

表 1: 実験に用いたデータの記述統計.

データセット	用途	文数	トークン数			トークンの異なり数		
			Ja	Fr	En	Ja	Fr	En
NTCIR	訓練データ	3M	110M		102M	169k		275k
	開発データ (pat-dev-2006-2007)	2,000	73k		67k	4k		5k
	評価データ (ntc9-je, 日英 T09)	2,000	74k		68k	5k		6k
	評価データ (ntc9-ej, 英日 T09)	2,000	74k		70k	5k		6k
	評価データ (ntc10-je, 日英 T10)	2,300	99k		92k	6k		7k
	評価データ (ntc10-ej, 英日 T10)	2,300	87k		80k	6k		6k
	単言語データ	-	27B		15B	9M		22M
WMT 2015	訓練データ	24M		726M	614M		2M	2M
	開発データ (newstest2012)	3,003		82k	73k		11k	10k
	評価データ (newstest2013, N13)	3,000		74k	70k		11k	9k
	評価データ (newstest2014, N14)	3,003		81k	71k		11k	10k
	単言語データ	-		2B	3B		4M	6M

番目の語 e_i の語彙翻訳確率であり，翻訳モデルの訓練データに単語アラインメントを適用して推定する．また，次式の通り，最尤の訳語のみを考慮することも考えられる [6].

$$P_{lmax}(e|f) = \frac{1}{I}\sum_{i=1}^{I}\log\Big(\max_j p(e_i|f_j)\Big) \qquad (2)$$

上の 2 つの式による文レベルの翻訳確率を，起点言語から目標言語 (順方向) と目標言語から起点言語 (逆方向) の各々について求め，リランキングに用いる．

2.4 単語事後確率に基づくスコア

PBSMT による翻訳候補のリランキングにおいて，翻訳候補集合における各トークンの出現確率に基づくスコアも有用である．我々は，式 (1) のスコアや NMT のモデルによる翻訳尤度を各翻訳候補の重みとし，重み付き出現頻度に基づく単語事後確率 [18] を計算する．

2.5 コンセンサススコア

各翻訳候補が，翻訳候補集合中の他の翻訳候補に平均的にどの程度類似しているかを定量化して用いる．翻訳候補間の類似度としては，文レベルの BLEU スコア [13] (sBLEU) や chrF++スコア [14] を用いる．

2.6 その他の素性

翻訳候補の長さ (単語ペナルティ(2.2 節) に相当) は，翻訳候補の品質を推定する上で有用な情報である [10]. 原文の長さとの差や比なども検討に値する．

いくつかの素性は，翻訳候補の出自によって極端に異なる値をとる．そのようなバイアスを制御するため，個々の翻訳候補が PBSMT と NMT のどちらによって生成されたものであるかを表す 2 値素性を用いる．

3 評価実験

各手法の性能を，日英・英日・仏英・英仏の 4 つの翻訳タスクを用いて評価した．

3.1 データ

日英・英日翻訳には，NTCIR の特許翻訳のデータ [1] を用いた．言語モデルを構築するため，NTCIR から提供されている単言語データも用いた．仏英・英仏翻訳には，WMT 2015[2] のニュース翻訳のデータを用いた．訓練データは，Europarl v7，10^9，news-commentary v10 からなる．単言語データとしては，News Crawl の 2007-2014 年分を用いた．データの種類と規模を表 1 に示す．

3.2 翻訳システム

PBSMT の実装としては，Moses[3] を用いた．日英・英日翻訳については mgiza[4]，仏英・英仏翻訳については fast_align[5] による単語アラインメントを適用した後，grow-diag-final-and ヒューリスティックスによってフレーズテーブルを構築した．言語モデルは，訓練データの目標言語側のみから 1 つ，それと単言語データを連結したものから 1 つ，4-gram モデルを Moses の lmplz を用いて構築した．語彙化並び替えモデルは，双方向の MSD モデルを用いた．各モデルの重み，および並び替え距離の上限値は，開発データと Moses の kbmira を用いて定めた．

NMT の実装としては，Nematus[6] をデフォルトの設定で用いた．BPE [16] を用いて，日英・英日翻訳については言語ごとに 5 万種の語彙を，仏英・英仏翻訳については両言語に共通の 5 万種の語彙を定めた．デコードの方向は L2R とした．翻訳モデルの学習の際，ミニバッチを 5,000 回処理するごとにモデルを保存し，学習の収束後，開発データに対する性能が最も良かった 4 つのモデルをアンサンブルに用いた．

[1] http://ntcir.nii.ac.jp/PatentMT-2/
[2] http://www.statmt.org/wmt15/
[3] https://github.com/moses-smt/mosesdecoder
[4] https://github.com/moses-smt/mgiza
[5] https://github.com/clab/fast_align
[6] https://github.com/EdinburghNLP/nematus

表 2: リランキングに用いた素性の一覧 (括弧内は素性の数).

素性	説明
L2R (5)	4 ベストの L2R の Nematus モデルによる翻訳尤度，およびそれらの幾何平均.
R2L (5)	4 ベストの R2L の Nematus モデルによる翻訳尤度，およびそれらの幾何平均.
PBFD (1)	PBFD のスコア.
LEX (4)	文レベルの翻訳確率．翻訳の順方向と逆方向の各々について，2 つの式 (1)，(2) に基づいて計算したもの.
LM (2)	言語モデルのスコア．Moses 用に構築した 2 種類の言語モデルの各々のスコアを使用.
WP (1)	単語ペナルティ.
WPP (2)	単語事後確率に基づくスコア．単語事後確率推定時の翻訳候補の重みとして，式 (1) に基づくものと Nematus による翻訳尤度の 2 種類を独立に使用.
MBR (2)	M+N におけるコンセンサススコア．*s*BLEU に基づくものと chrF++に基づくものの 2 種類.
LEN (2)	原文と翻訳候補の長さの差分およびその絶対値.
SYS (1)	翻訳候補を生成したシステムに関するフラグ．M に含まれていれば 1．N に含まれていれば 0.

表 3: 各手法によって生成・選択された訳文の BLEU スコア.

手法	リランキング対象	日英		英日		仏英		英仏	
		T09	T10	T09	T10	N13	N14	N13	N14
PBSMT: Moses	-	31.3	32.7	34.7	35.9	31.4	39.9	30.6	39.1
NMT: Nematus	-	41.9	41.6	44.8	45.4	30.8	34.0	30.6	35.8
既存手法 (a): Jane	M+N (n=1)	41.5	41.6	44.9	45.8	32.0	40.0	30.8	39.5
既存手法 (a): Jane	M+N (n=100)	39.0	40.2	41.6	42.7	32.1	40.0	31.0	39.9
既存手法 (b): R_{nmt}	M	33.3	34.1	36.8	38.3	33.6	41.4	32.4	40.5
既存手法 (c): R_{pbsmt}	N	42.5	43.1	46.1	46.7	32.5	34.7	31.4	36.1
提案手法	M	33.5	34.2	36.7	38.5	33.6	41.4	32.4	40.4
提案手法	N	43.0	43.8	46.9	47.5	33.9	38.0	32.3	38.8
提案手法	M+N	43.0	43.9	47.1	47.7	34.2	41.6	32.6	40.8

3.3 リランキングシステム

既存手法 (a) のうち Confusion Network に基づくものを，Jane [4] および PBSMT 用に構築した言語モデルを用いて実現した．Moses と Nematus の各々の翻訳候補の個数として，1 個と 100 個を比較した.

既存手法 (b) に相当するシステム R_{nmt} では，Moses による 100 ベストの翻訳候補 (M) を，Moses の全ての素性，Nematus の 4 ベストの L2R モデル，4 ベストの R2L モデルの尤度を用いてリランキングした．一方，既存手法 (c) に相当する R_{pbsmt} は，Nematus のビーム幅 [7] を 100 として得た 100 ベストの翻訳候補 (N) を，文献 [19] と同様に，Nematus の 4 ベストの L2R モデルの尤度，Moses のフレーズテーブルに基づく PBFD のスコア，単語ペナルティを用いてリランキングした.

既存手法 (d) も実装・評価したが，全ての翻訳タスクにおいて，単体の翻訳システムよりも性能が低かったため，以下では言及しない.

我々の提案手法は，開発データに対する M と N をあわせたもの (以下，M+N)，表 2 に示す 25 種類の素性 [8]，および Moses の kbmira を用いて実現した.

[7]翻訳候補の品質の低下を避けるため，デコード時に参照する尤度を訳文の長さ (トークン数) で正規化した [7].

[8]PBFD の結果に基づいて計算した歪みスコア，語彙化並び替えモデルのスコア，翻訳確率，フレーズペナルティ，および PBSMT の尤度も利用できる (2.2 節) が，今回の実験においてはいずれも性能を劣化させたため，以下では言及しない.

3.4 評価結果

各手法によって生成・選択された訳文の BLEU スコア [13] を表 3 に示す．最初の 2 行が示す通り，PBSMT と NMT の優劣は翻訳タスクによって異なる．日英・英日翻訳については，長距離の並び替えをうまく扱える NMT が，PBSMT よりも 8.9〜10.6 ポイント高い BLEU スコアを達成した．一方，仏英・英仏翻訳については，PBSMT の方が性能が良かった.

既存手法 (a) の性能は，PBSMT よりも常に高かった．また，日英翻訳以外の 3 つの翻訳タスクにおいて，NMT の性能も改善した．ただし，100 ベストの翻訳候補を用いた場合，日英・英日翻訳の性能は 1.4〜3.3 ポイント劣化し，NMT を下回った．これは，PBSMT による翻訳候補が極端に低品質であるためと考えられる.

既存手法 (b) は PBSMT よりも 1.4〜2.4 ポイント，既存手法 (c) は NMT よりも 0.3〜1.7 ポイント高い BLEU スコアを示した．既存手法の中では，日英・英日翻訳では手法 (c) が，仏英・英仏翻訳では手法 (b) が最も高い性能を示した．PBSMT と NMT の優劣は，そのまま手法 (b) と (c) の優劣に継承されている.

我々の提案手法は，M のみを対象とした場合は既存手法 (b) と同程度の性能であったが，N のみを対象とした場合は既存手法 (c) よりも 0.5〜3.3 ポイント高い BLEU スコアを達成した．さらに，M+N を対象とすることにより，4 つの翻訳タスク，8 種類の評価データの全てに

表 4: 一部の素性を除外して学習した提案手法の性能.

除外した素性	日英, T09	英日, T09	仏英, N13	英仏, N13
なし	43.0	47.1	34.2	32.6
L2R	43.0	47.1	**34.1**	**32.5**
R2L	**42.6**	**46.6**	**33.7**	**32.1**
PBFD	43.2	**47.0**	34.3	**32.3**
LEX	43.0	47.1	**34.1**	**32.4**
LM	43.1	47.1	**34.0**	**32.2**
WP	43.1	47.1	**34.1**	32.6
WPP	43.1	47.1	**34.1**	**32.4**
MBR	43.1	**46.9**	**33.9**	**32.4**
LEN	43.1	47.1	**32.7**	**32.0**
SYS	43.1	47.1	**34.1**	32.6
L2R+R2L	**42.5**	**46.7**	**33.7**	**32.0**
PBFD+LEX	43.1	**47.0**	34.3	**32.5**
WPP+MBR	43.2	**46.9**	**34.0**	**32.4**
WP+LEN	43.0	**46.7**	**32.8**	**31.9**
L2R+R2L+ PBFD+LEX	**42.4**	**46.2**	**33.6**	**31.8**

おいて，最も高い性能を示した．PBSMT と NMT の良い方のシステムに対する BLEU スコアの改善幅は 1.1 〜2.8 ポイントであった．相対的に品質が低い翻訳候補を加えた場合でも性能が劣化しておらず，提案手法の頑健さが示された．

表 2 の素性の一部を除外して学習した提案手法の，日英・英日翻訳の T09，仏英・英仏翻訳の N13 における BLEU スコアを表 4 に示す．各素性の影響の多寡は翻訳タスクによって異なっていた．意外なことに，仏英翻訳における LEN (-1.5 ポイント) を除き，素性を 1 つ除外しても，BLEU スコアは最大 0.5 ポイントしか下がらなかった．NMT に基づく素性 (L2R+R2L) と翻訳確率に関する素性 (PBFD+LEX) は計算コストが非常に高いが，これらを除外しても，BLEU スコアは最大 0.9 ポイントしか下がらなかった．NMT の尤度については，L2R よりも R2L の方が有用であった [15]．L2R のモデルで生成した翻訳候補を用いたためであろう．

4 おわりに

本稿では，PBSMT と NMT の n ベストの翻訳候補を合わせてリランキングすることにより，両者の優劣に関わらず性能を改善できることを示した．これは 2 節で述べた，リランキングの対象とする翻訳候補の多様性 [5] に依拠する．翻訳候補の多様性，品質に関する分析結果については，本稿の詳細版 [9] を参照されたい．

今後は，より小規模な対訳データしか存在しない状況における提案手法の有用性を検証する予定である．

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た．

参考文献

[1] S. Bangalore, G. Bordel, and G. Riccardi. Computing consensus translation from multiple machine translation systems. In *Proc. of the ASRU*, pp. 351–354, 2001.

[2] L. Bentivogli, A. Bisazza, M. Cettolo, and M. Federico. Neural versus phrase-based machine translation quality: a case study. In *Proc. of EMNLP*, pp. 257–267, 2016.

[3] J. Du and A. Way. Neural pre-translation for hybrid machine translation. In *Proc. of MT Summit*, pp. 27–40, 2017.

[4] M. Freitag, M. Huck, and H. Ney. Jane: Open source machine translation system combination. In *Proc. of EACL*, pp. 29–32, 2014.

[5] K. Gimpel, D. Batra, C. Dyer, and G. Shakhnarovich. A systematic exploration of diversity in machine translation. In *Proc. of EMNLP*, pp. 1100–1111, 2013.

[6] A. S. Hildebrand and S. Vogel. Combination of machine translation systems via hypothesis selection from combined n-best lists. In *Proc. of AMTA*, pp. 254–261, 2008.

[7] P. Koehn and R. Knowles. Six challenges for neural machine translation. In *Proc. of the First Workshop on Neural Machine Translation*, pp. 28–39, 2017.

[8] H.-S. Le, A. Allauzen, and F. Yvon. Continuous space translation models with neural networks. In *Proc. of NAACL-HLT*, pp. 39–48, 2012.

[9] B. Marie and A. Fujita. A smorgasbord of features to combine phrase-based and neural machine translation. In *Proc. of AMTA*, 2018. (to appear).

[10] D. S. Munteanu and D. Marcu. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, Vol. 31, No. 4, pp. 477–504, 2005.

[11] J. Niehues, E. Cho, T.-L. Ha, and A. Waibel. Pre-translation for neural machine translation. In *Proc. of COLING*, pp. 1828–1836, 2016.

[12] F. J. Och, D. Gildea, S. Khudanpur, A. Sarkar, K. Yamada, A. Fraser, S. Kumar, L. Shen, D. Smith, K. Eng, V. Jain, Z. Jin, and D. Radev. A smorgasbord of features for statistical machine translation. In *Proc. of HLT-NAACL*, pp. 161–168, 2004.

[13] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proc. of ACL*, pp. 311–318, 2002.

[14] M. Popović. chrF++: words helping character n-grams. In *Proc. of WMT*, pp. 612–618, 2017.

[15] R. Sennrich, A. Birch, A. Currey, U. Germann, B. Haddow, K. Heafield, A. V. Miceli Barone, and P. Williams. The University of Edinburgh's neural MT systems for WMT17. In *Proc. of WMT*, pp. 389–399, 2017.

[16] R. Sennrich, B. Haddow, and A. Birch. Neural machine translation of rare words with subword units. In *Proc. of ACL*, pp. 1715–1725, 2016.

[17] A. Toral and V. M. Sánchez-Cartagena. A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. In *Proc. of EACL*, pp. 1063–1073, 2017.

[18] N. Ueffing and H. Ney. Word-level confidence estimation for machine translation. *Computational Linguistics*, Vol. 33, No. 1, pp. 9–40, 2007.

[19] J. Zhang, M. Utiyama, E. Sumita, G. Neubig, and S. Nakamura. Improving neural machine translation through phrase-based forced decoding. In *Proc. of IJCNLP*, pp. 152–162, 2017.