翻訳の品質評価に基づく動的な混成サンプリングによるNMT の双方向反復逆翻訳手法の改善

森田 知熙　秋葉 友良　塚田 元
豊橋技術科学大学
{morita.tomohiro.al,akiba.tomoyoshi.tk,tsukada.hajime.hl}@tut.jp

概要

本稿では，ニューラル機械翻訳における Iterative Back-Translation(IBT) を用いたドメイン適応手法のための新しいサンプリング手法を提案する．従来の IBT では，各段階の翻訳モデルの性能によらず，単言語コーパスのサンプリング手法は一様であった．本手法では，疑似原文の品質評価に基づき動的に逆翻訳時のサンプリング手法を決定する手法を提案する．WMT14 ドメインから TED ドメインへの英独，独英方向のドメイン適応実験の結果，従来のサンプリング手法による IBT に比べ，最大で独英方向では 0.80，独英方向では 0.74 ポイント BLEU が向上した．

1　はじめに

近年の機械翻訳研究は，ニューラル機械翻訳 (NMT) が主流となっている．NMT モデルの翻訳性能は，対訳コーパスのサイズと質に大きく影響を受けるが，対訳コーパスの収集，構築は困難であり，対訳コーパスが存在しないドメインの NMT モデルの学習も難しい．一方で，対訳になっていない単言語コーパスは構築が容易である．そのため，単言語コーパスを用いて NMT モデルの翻訳性能を向上させる手法が提案されている．Sennrich ら [1] は目的言語側の単言語コーパスを逆翻訳して擬似対訳コーパスを生成し，これを既存の対訳コーパスと組み合わせて学習を行う手法を提案した．

Hoang ら [2]，Zhang[3] らは，この手法を拡張し，原言語，目的言語の両方の単言語コーパスを用いて，両言語間で繰り返し逆翻訳による擬似対訳コーパスの生成と NMT モデルの学習を行う Iterative Back-Translation(IBT) を提案した．藤澤ら [4]，Jin ら [5]，森田ら [6] はこの手法をドメイン適応手法として用い，有効性を示している．

本研究では，IBT を用いたドメイン適応において，各文ごとに動的にサンプリング手法を変更することで，最終的な NMT モデルの性能向上を図る．具体的には，疑似原文の品質評価に基づくサンプリング手法の決定を行うことでより翻訳モデルの性能に即した有用な疑似原文を生成する手法を提案する．

WMT14 ドメインから TED ドメインへのドメイン適応において，英独，独英方向の実験を行った結果，品質評価に基づく手法は，ベースラインモデルから最大で 0.80 ポイント BLEU が向上した．

2　関連研究

単言語コーパスを活用した半教師あり学習手法として，Sennrich ら [1] は目的言語側の単言語コーパスを逆翻訳し，疑似原文を生成することで擬似対訳コーパスを作る手法を提案した．今村ら [7],Edunov ら [8] はこの手法を発展させ，逆翻訳時にランダムサンプリングを用いることで疑似原文の多様性が増加し，最終的な翻訳モデルの性能が向上することを報告した．Hoang ら [2] や Zhang[3] らは，Sennrich ら [1] の手法を両方向に繰り返し用いる Iterative Back-Translation(IBT) を提案した．藤澤ら [4]，Jin ら [5]，森田ら [6] はこの手法をドメイン適応手法として用い，その有効性を評価した．森田ら [9] は，IBT におけるモデルの更新時に Fine-Tuning を活用する実験を行い，Fine-Tuning により各方向の翻訳モデルの性能を有効活用できることを示した．疑似原文の品質を向上させる本研究同様のアプローチとして，Wei ら [10] は Domain-Repaired モデルを IBT に組み込むことで，疑似原文のノイズを除去し，品質を向上させる手法を提案した．Dou ら [11] は，IBT において，Round-trip BLEU(R-BLEU) や Sentence BERT(S-BERT) で得た各文の分散表現のコサイン類似度により擬似対訳コーパスのフィルタリングと学習時の重み付けを行う手法を提案した．この手

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

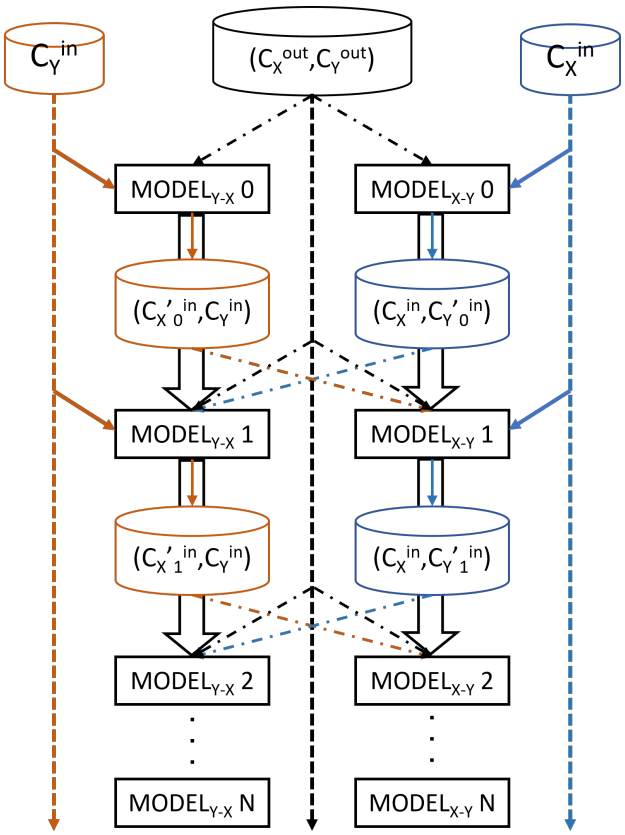

図 1 Iterative Back-Translation の流れ

法は，本研究と同様に疑似原文の品質評価により，スコアリングと学習時の重み付けを行っているが，サンプリング手法そのもの検討は行われておらず，コーパスの多様性を増す手法の検討は行われていない．

3 Iterative Back-Translation

IBT では，図 1 のように対訳コーパス (C_X^{out}, C_Y^{out}) と対訳でない単言語コーパス C_X^{in}, C_Y^{in} を用い，反復的に逆翻訳と学習を行う．原言語を X，目的言語を Y とする．また，X から Y への翻訳を X-Y とし，Y から X への翻訳を Y-X とする．IBT の流れを以下に示す．

1 対訳コーパス (C_X^{out}, C_Y^{out}) から両方向の翻訳モデル $(Model_{X-Y}0, Model_{Y-X}0)$ を学習する．
2 以下の手順で $Model_{X-Y}(i)$ を更新する．
  2.1 $Model_{Y-X}i$ で単言語コーパス C_Y^{in} を翻訳し，疑似対訳コーパス $(C_X^{in\prime}, C_Y^{in})$ を得る．
  2.2 疑似対訳コーパス $(C_X^{in\prime}, C_Y^{in})$ と対訳コーパス (C_X^{out}, C_Y^{out}) を結合し，$Model_{X-Y}(i)$ から fine-tuning し，$Model_{X-Y}(i+1)$ を学習する．
3 以下の手順で $Model_{Y-X}i$ を更新する．
  3.1 $Model_{X-Y}i$ で単言語コーパス C_X^{in} を翻訳し，疑似対訳コーパス $(C_Y^{in\prime}, C_X^{in})$ を得る．
  3.2 疑似対訳コーパス $(C_Y^{in\prime}, C_X^{in})$ と対訳コーパス (C_X^{out}, C_Y^{out}) を混合し，$Model_{Y-X}i$ から fine-tuning し $Model_{Y-X}(i+1)$ を学習する．
4 i ← i+1 としてステップ 2 に戻る．

4 提案法

4.1 動的混成サンプリング

Ednov[8] らの研究結果から，逆翻訳器が正しく単言語コーパスを翻訳できる場合はランダムサンプリングによる逆翻訳が有効に働き，そうでない場合にはビームサーチが有効に働くことが予想できる．提案手法では，疑似原文の品質を評価することにより，動的に単言語コーパスのサンプリング手法を決定する．疑似原文は，もとの単言語コーパスと異なる言語であるため，そのままでは品質評価が困難である．そのため，IBT の各段階で翻訳モデルが生成した疑似原文を，反対方向の翻訳モデルで再度翻訳（往復翻訳）する．往復翻訳された文は，元々の単言語コーパスと同一の言語の文になるため，BLEU や文の分散表現を用いた品質評価が可能となる．

品質評価のために，まずは全ての単言語コーパスをビームサーチにより逆翻訳し，同様に往復翻訳を行う．次に，往復翻訳文品質評価尺度によるスコリングを行い，スコアが予め定めたしきい値を超えた文は，ランダムサンプリングに翻訳した文に置き換える．

本手法の概略図を図 2 に示す．本手法を用いて $Model_{X-Y}(n)$,$Model_{Y-X}(n)$ から $Model_{X-Y}(n+1)$ を学習する場合の流れは以下のようになる．

1 $Model_{Y-X}i$ で単言語コーパス C_Y^{in} をビームサーチにより翻訳し，逆翻訳文 $C_X^{in\prime}$ を得る．
2 $Model_{X-Y}i$ で逆翻訳文 $C_X^{in\prime}$ をビームサーチにより翻訳し，往復翻訳文 $C_Y^{in\prime\prime}$ を得る．
3 単言語コーパス C_Y^{in} を参照訳として往復翻訳文 $C_Y^{in\prime\prime}$ の翻訳品質評価スコアを文ごとに算出し，各文のスコアとする．
4 スコアがしきい値 α より高い文を，$Model_{Y-X}i$ で単言語コーパス C_Y^{in} をランダムサンプリングにより翻訳した文で置き換え，疑似原文 $\hat{C}_X^{in}$ を得る．
5 単言語コーパス C_Y^{in} と最終的な疑似原文 $\hat{C}_X^{in}$ を

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

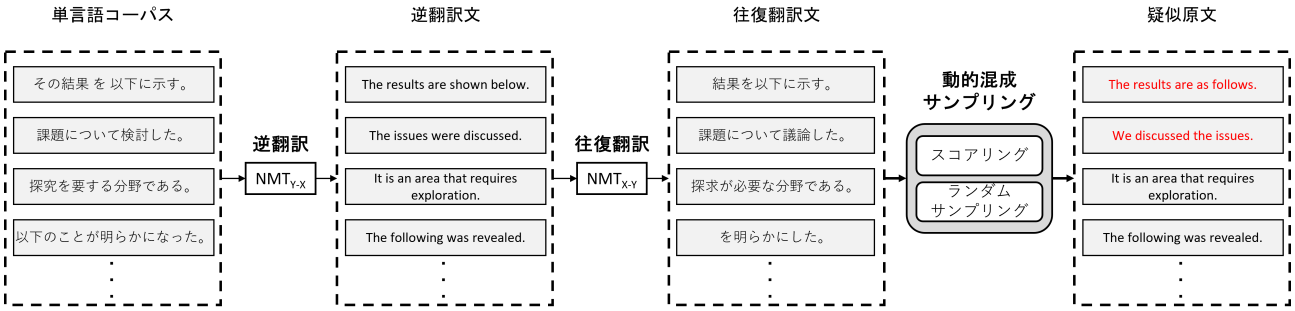

図 2　動的混成サンプリングの流れ

擬似対訳コーパスとし，$Model_{X-Y}(n+1)$ を学習する．

$Model_{Y-X}(n+1)$ も同様の手順により学習可能である．

本手法では，サンプリング手法を決定するために各文をスコアリングするための品質評価尺度が必要となる．本研究では，往復翻訳文の品質評価に R-BLEU と S-BERT により抽出した文の分散表現のコサイン類似度を用いた．それぞれの詳細を以下に示す．

4.1.1　Round-Trip BLEU(R-BLEU)

BLEU[12] は，機械翻訳モデルの性能評価に広く使われている評価尺度であり，参照訳と翻訳モデルが生成した文の単語の一致数をもとに算出される．本手法では，元々の単言語コーパスと，往復翻訳文の BLEU を疑似原文のスコアとする．

4.1.2　Sentence BERT(S-BERT)

BERT[13] とは，大規模な単言語コーパスにより学習された汎用な言語表現モデルである．また，BERT をファインチューニングし，文の分散表現の獲得に特化させた S-BERT[14] が提案されている．本手法では，S-BERT により元々の単言語コーパスと往復翻訳文それぞれの文の分散表現を獲得し，それらのコサイン類似度を疑似原文のスコアとして用いる．

5　実験

5.1　データセット

本実験では，ドメイン外対訳コーパスとして WMT14 コーパスを，ドメイン内単言語コーパスとして TED-talks を使用した．翻訳タスクは，英語-ドイツ語とドイツ語-英語の双方向とした．

TED コーパスは，約 15 万文を半分に分割し，先頭の 7 万文を英語，後ろの 7 万文をドイツ語の単言語コーパスとみなすことで 2 言語間で単言語コーパスが対訳とならないようにした．

前処理として，すべてのデータに対し NFKC 正規化し，Moses[15] の truecaser により表記の統一を行った．また，Moses tokenizer により単語分割し，Sentence Piece[16] によりサブワード化を行った．Sentence Piece のモデルの学習には，ドメイン外対訳コーパスとドメイン内単言語コーパスを結合したものを用いた．

5.2　実験設定

翻訳システムは，OpenNMT-py[17] の Transformer を使用した．エンコーダ，デコーダ共に 6 層，隠れ層の次元を 512 とした．各モデルの性能評価には BLEU を用いた．学習ステップは，10 エポック分となるように設定し，各エポックごとにチェックポイントを保存した．開発データに対する BLEU が最も高いチェックポイントのモデルをテストデータでの性能評価と単言語コーパスの逆翻訳に使用した．

S-BERT は，Sentence Transformers[14] を使用し，事前学習済みのモデルは"all-MiniLM-L6-v2"を使用した．動的混成サンプリングのしきい値は，R-BLEU を用いたモデルでは 0.65，S-BERT を用いたモデルでは 0.95 とした．これらの値は，いくつかのしきい値を設定して実験をおこなかった中で，開発データに対する BLEU が最も高くなるものを使用した．

提案手法の効果を評価するために，以下の手法で性能を比較した．

ベストサンプリング　通常の beam search により IBT を行う．

ランダムサンプリング　IBT における単言語コーパスの逆翻訳時に random sampling により文を生成し逆翻訳を行う．

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

表 1 各モデルの IBT における BLEU の最大値

	TED	
Model	En-De	De-En
ベストサンプリング	31.40	37.30
ランダムサンプリング	26.19	30.54
混成サンプリング (5:5)	31.32	36.82
混成サンプリング (7:3)	31.58	37.06
混成サンプリング (9:1)	31.57	37.08
動的混成サンプリング (R-BLEU)	32.11	38.04
動的混成サンプリング (S-BERT)	32.20	37.97
教師あり学習	31.19	39.07

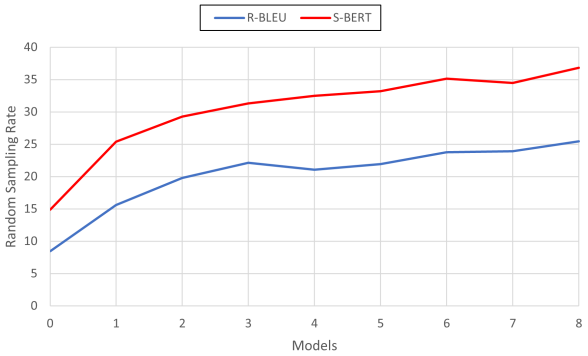

図 3 独英翻訳モデルの各ステップにおける動的混成サンプリングのランダムサンプリングの割合

混成サンプリング [9] 単言語コーパスを事前に分割し，上記の二つの方法を組み合わせて疑似対訳コーパスを作成する．今回の実験では，単言語コーパスを予め 5:5, 9:1, 7:3 の割合で分割し，片方をベストサンプリング，他方をランダムサンプリングで逆翻訳する．なお，IBT の各ステップで単言語コーパスの分割の仕方を変更する．

動的混成サンプリング 提案手法により疑似原文の品質を評価し，文単位で動的にサンプリング手法を変更する．品質評価尺度として，R-BELU, S-BERT の 2 つの尺度を用い，それぞれのスコアがしきい値を超えた文をランダムサンプリングで逆翻訳する．それ以外の文は，ベストサンプリングで逆翻訳された文をそのまま学習に用いる．

5.3 実験結果

表 1 に各手法で Model 9 まで IBT を行った中で，TED の開発データに対する BLEU が最も高かったモデルのテストデータに対する BLEU を示す．混成サンプリングではベストサンプリングによる IBT を行ったベースラインから翻訳性能の改善は見られなかった．また，混成サンプリングの結果を比較すると，ベストサンプリングとランダムサンプリングの割合が 5:5 のモデルが最も BLEU が低く，両方向でベストサンプリングを下回っているのに対し，混合割合が 9:1 のモデルでは英独方向ではベースラインを+0.17 上回っている．このことから，本データセットではベストサンプリングが有効であると考えられる．一方で，動的混成サンプリングを用いたモデルは，R-BLEU, S-BERT 共に BLEU が向上し，ベースラインから英独方向では最大で+0.80, 独英方向では+0.74 ポイント BLEU が向上した．これは，翻訳モデルが高品質な翻訳を行うことができる単言語コーパスに関してはランダムサンプリングが有効であり，逆に正しい翻訳を学習できていない単言語コーパスにはベストサンプリングによる逆翻訳が適しているという予想と一致する．また，R-BLEU, S-BERT のどちらを使用した場合もベースラインから翻訳性能は向上しているが，それぞれの品質評価尺度での文ごとのスコアの傾向は必ずしも一致しているとは限らないため，２つのスコアを組み合わせることでより翻訳精度が向上する可能性がある．図 3 は，動的混成サンプリングの各ステップにおけるサンプリング手法の割合の推移を表している．学習が進むにつれ，ランダムサンプリングの割合が増加していることから，提案手法が IBT における翻訳モデルの性能の変化を活用できていることがわかる．

6 結論

本研究では，IBT によるドメイン適応手法の性能を改善するために，往復翻訳による品質評価に基づいたサンプリング手法を提案した．実験の結果，動的混成サンプリングを用いた 2 つのモデルは共にベストサンプリングによる IBT を行ったベースラインを上回った．従来の手法である混成サンプリングでは，翻訳性能が向上しなかったため，本手法は，ランダムサンプリングを用いた従来の手法では性能改善が難しいドメインでも有効に働くことことが確認できた．今後の課題としては，現状の動的混成サンプリングは人手によりしきい値を定める必要があるため，パラメータの探索の必要がないサンプリング手法を検討したい．

謝辞 本研究は JSPS 科研費 19K11980 および 18H01062 の助成を受けた．

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

参考文献

[1] Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In **Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 86–96, Berlin, Germany, August 2016. Association for Computational Linguistics.

[2] Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. Iterative back-translation for neural machine translation. In **Proceedings of the 2nd Workshop on Neural Machine Translation and Generation**, pp. 18–24, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[3] Zhirui Zhang, Shujie Liu, Mu Li, Ming Zhou, and Enhong Chen. Joint training for neural machine translation models with monolingual data. **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 32, No. 1, Apr. 2018.

[4] 藤澤兼太, 秋葉友良, 塚田元. ニューラル機械翻訳における双方向反復的教師なし適応の改善. 言語処理学会 第 26 回年次大会 発表論文集, pp. 744–747. Association for Natural Language Processing, March 2020.

[5] Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. A simple baseline to semi-supervised domain adaptation for machine translation, 2020.

[6] 森田知熙, 秋葉友良, 塚田元. 双方向の逆翻訳を利用したニューラル機械翻訳の教師なし適応の検討. 言語処理学会 第 25 回年次大会 発表論文集, pp. 1451–1454. Information Processing Society of Japan, March 2019.

[7] 今村賢治, 藤田篤, 隅田英一郎. サンプリング生成に基づく複数逆翻訳を用いたニューラル機械翻訳. 人工知能学会論文誌, Vol. 35, No. 3, pp. A–JA9_1–9, 2020.

[8] Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. Understanding back-translation at scale. In **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**, pp. 489–500, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[9] 森田知熙, 秋葉友良, 塚田元. Fine-tuninng と混成的な逆翻訳サンプリングに基づく nmt の双方向反復的教師なし適応の改善. 言語処理学会 第 27 回年次大会 発表論文集, pp. 1669–1674. Information Processing Society of Japan, March 2021.

[10] Hao-Ran Wei, Zhirui Zhang, Boxing Chen, and Weihua Luo. Iterative domain-repaired back-translation. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 5884–5893, Online, November 2020. Association for Computational Linguistics.

[11] Zi-Yi Dou, Antonios Anastasopoulos, and Graham Neubig. Dynamic data selection and weighting for iterative back-translation. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 5894–5904, Online, November 2020. Association for Computational Linguistics.

[12] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In **Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics**, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[14] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics.

[15] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 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**, pp. 177–180, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[16] Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations**, pp. 66–71, Brussels, Belgium, November 2018. Association for Computational Linguistics.

[17] Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. OpenNMT: Open-source toolkit for neural machine translation. In **Proceedings of ACL 2017, System Demonstrations**, pp. 67–72, Vancouver, Canada, July 2017. Association for Computational Linguistics.

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

A　付録

動的混成サンプリングの各ステップにおける英独方向のサンプリング手法の割合の推移を以下に示す.

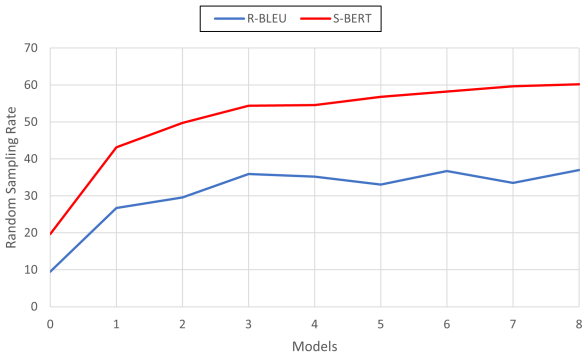

図 4　英独翻訳モデルの各ステップにおける動的混成サンプリングのランダムサンプリングの割合

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).