言語横断な言語モデルによる原言語情報を活用した自動機械翻訳評価

高橋 洸丞[1,a)]　須藤 克仁[1,2,b)]　中村 哲[1,c)]

概要：本研究では原言語文と参照訳文から翻訳文がどれだけ正しいかを推定する自動評価手法を検討する. 既存の自動翻訳評価手法では, 1 対 1 の参照訳文と翻訳文のペアから翻訳文を評価する手法が主流だが, 1 対 1 ペアの比較では, 翻訳文が参照訳文と一見異なるが正しい訳文である際に評価性能が下がりやすい. この問題は, マルチリファレンスと呼ばれる参照訳文を複数用意することで解決可能だが, 各原言語文に対して参照訳文を複数作成する必要がありコストが高い. そこで本研究では, 原言語文も参照訳文の一つとみなして, 原言語文を評価に用いることで前述の問題に対処する. 言語横断な言語モデルを用いて入力文を分散表現とし, 最終的な評価値を多層パーセプトロンを通して出力する回帰モデルによる評価手法を検討した.

1. はじめに

機械翻訳の自動評価手法は, 人手評価と比較して低コストで素早く評価に変動がないことから, 機械翻訳システム研究で広く用いられており, 人手評価との高い相関が得られるよう様々な手法が研究されてきた. 既存の多くの自動評価手法は, 現在に至るまで標準的に利用されている BLEU[13] を始めとして, 参照訳文と呼ばれる翻訳後の正解訳文と機械翻訳による翻訳文を 1 対 1 で比較し, 翻訳文が参照訳文にどれだけ近いかを評価値として算出する. BLEU では単語 n-gram の表層的な一致率に基づいて文の近さを評価するが, 同じ意味の文でも同意語による置換や語の並びによって評価値が大きく変動してしまう. そこで, 同意語辞書を用いて同意語のマッチングを緩和する METEOR[1] や, より単語の意味に着目し, 単語分散表現上の距離に基づいた Word Mover's Distance[9] や, 単語分散表現のファジーマッチングにより文意の近さを評価する bleu2vec[18] がある. また, 単語ではなく直接文の分散表現を用いた評価手法 RUSE[6], BERT ベースの RUSE[20] はより大域的な文意を比較できるため, 高い評価性能を達成している.

ここで前述の自動評価手法は, 参照訳文と翻訳文の 1 対 1 のペアから評価値を算出するシングルリファレンスである. 原言語文に対して正解となりうる参照訳文を考えると, 文のスタイル, 使用単語, 単語の並びなどの違いにより本来複数の参照訳文の候補が存在するため, 複数の参照訳文と照らし合わせて翻訳文を評価する方がより柔軟な評価が可能となる. そして, このマルチリファレンスと呼ばれる参照訳文を複数文用いた評価手法として, [12], [14] がある. BLEU[13], NIST[4], HTER[17] をマルチリファレンスでの評価値計算を可能にし, マルチリファレンス下での評価実験により人手評価との相関性が向上することが示されている. しかし, マルチリファレンスを用いた評価は現在主流ではない. その原因が, マルチリファレンスは正しい参照訳を大量に集めることが難しいことや, 人手での収集はコストが高く, 機械翻訳や言語モデルによるパラフレージング [7], [15] は正解訳としての信頼性に欠けてしまうという点にある.

そこで本研究では, 原言語文を擬似的な参照訳文として用いる機械翻訳の自動評価手法を検討する. 翻訳の入力である原言語文は参照訳文と同意であるため, 言語の差はあるものの原言語文を同じ正解訳として扱えると考えた. そして, 言語の差から直接的に原言語文を用いて翻訳文の評価を行うのは難しいため, 原言語文と目的言語文を同一ベクトル空間で表現できる言語横断型な言語モデル XLM (Crosslingual Language Model)[10] を用いることにした. また, 翻訳文と参照訳文を比較し評価値を算出する基準となるシステムは, 現在文レベルの自動評価手法として最も人手評価と相関性の高い BERT(Bidirectional Encoder Representations from Transformers)[8] ベースの回帰モデル [20] を参考に, XLM をベースとした回帰モデルとした.

[1] 奈良先端科学技術大学院大学
NAIST, Takayama-cho 8916-5, Ikoma-shi, Nara 630–0192, Japan
[2] 科学技術振興機構さきがけ
PRESTO, Japan Science and Technology Agency
[a)] takahashi.kosuke.th0@is.naist.jp
[b)] sudoh@is.naist.jp
[c)] s-nakamura@is.naist.jp

2. 関連研究

2.1 SentBLEU

SentBLEU[2] は Moses[*1]のツールキットに含まれる評価手法で, BLEU(Bilingual Evaluation Understudy)[13] にスムージング処理を加えた文レベルの自動評価手法である. BLEU は現在様々な分野で基準とされている自動評価で, 参照訳文と翻訳文さえあれば評価値を算出できる手法である. 評価値は一般的に 1〜4 の単語 n-gram による参照訳文と翻訳文の一致率であるが, $n = 4$ のときなど大きい n で参照訳文と翻訳文の単語一致がない場合, 評価値が 0 になってしまう問題があるため, スムージング処理がされる.

2.2 RUSE

RUSE (Regressor Using Sentence Embeddings)[6] は WMT-2018 Metrics Shared Task[11] において, 文単位の評価が最も人手評価と高い相関を示した自動評価手法である. RUSE の大きな特徴は, 18 年度以前まで WMT のタスクで主流であった単語 n-gram による特徴を用いるのではなく, 事前に大規模なコーパスで学習した符号化器による文ベクトルを用いる点にある. 単語 n-gram に比べて一文を分散表現として扱うことで, より大域的な情報を考慮することや, 同意な単語や句の置換にも強い. 評価値を算出するモデルは, 符号化器で得られた文ベクトルを入力とした, 多層パーセプトロン (MLP) に基づく回帰モデルで, 人手評価値 [5] と相関性が高くなるように学習する.

2.3 BERT ベースの RUSE

BERT ベースの RUSE[20] は, RUSE の符号化器を BERT[8] に置き換え, 符号化器も MLP と共に学習する回帰モデルで, WMT-2018 Metrics Shared Task データにおいて最高性能を達成した. ここで BERT とは, 質問応答や含意関係認識などの様々な共通タスクで高い性能を記録した, Transformer[19] による双方向言語モデルである.

RUSE では事前学習された文符号化の有効性が示されたが, BERT をベースとしたモデルでは, 事前学習された文符号化がより高性能なモデルに置き換えられたことで, 評価性能の向上に繋がったと考えられる.

3. XLM による機械翻訳評価

本研究では, 15 言語の文意を識別する XNLI15[3] タスクで, BERT よりも高い性能を発揮した言語横断型な言語モデル XLM[10] を用いて, 自動評価における原言語情報の有用性を検証した. XLM は, 非言語依存なサブワード BPE(Byte Pair Encoding)[16] により, 同じ文字を共有する言語間の埋め込み空間のアライメントを強化している. BERT と異なる点は, 学習時に同意な文を二言語用意し一つの言語でマスキングされた部分をもう一つの言語の文から予測する学習ステップがあることと, 言語埋め込み表現 (language embeddings) と呼ばれる言語の ID を加えている点である (**図 1**). 原言語と目的言語の二種類の言語を同時に扱うとき, 単一言語のみに対応する言語モデルを文の符号化器として扱うと, 原言語と目的言語間に言語差によるベクトルの次元のずれが生じてしまい, 上手く原言語情報を活かせない. そこで原言語文を参照訳文と同様に扱うために, 言語による文ベクトルの次元を揃えられる言語横断型の XLM を符号化器とすることにした.

従来手法である BERT ベースの RUSE, 提案手法 1 と提案手法 2 の構成をそれぞれ**図 2(a)**, **図 2(b)**, **図 2(c)** に示す. 提案手法 1 と提案手法 2 の違いは XLM への入力方法が異なる点である. 提案手法 1 では RUSE[6] や BERT ベースの RUSE[20] の符号化器への文入力の方式に習って, 原言語文-翻訳文, 参照訳文-翻訳文の二つのペアを独立に XLM(XNLI 符号化モデル) へ入力し, 得られた分散表現を結合して MLP で人手評価値を回帰タスクとして推定する. 提案手法 2 では原言語文-参照訳文-翻訳文の文ペアをまとめた一つの入力にして分散表現を XLM から得る. その後の分散表現は提案手法 1 と同じく, MLP で回帰タスクとして入力し学習を行った. また, 学習時の誤差関数は平均二乗誤差 (MSE) とし, MLP から XLM への誤差逆伝搬により MLP と XLM 符号化器を共に学習した.

4. 評価実験

XLM により, 原言語文を評価に用いる提案手法の有効性を評価するために, 以下の実験設定で評価実験を行った.

4.1 実験設定

評価実験は WMT-2017 Metrics Shared Task[2] のコーパスを用いた. コーパス内の英語への翻訳タスクデータから, XLM が事前学習済みの言語対であるドイツ語 (de)-英語 (en), ロシア語 (ru)-英語, トルコ語 (tr)-英語を提案手法の XLM を用いた回帰モデルで使用した. WMT-2015, 2016 のデータを訓練データ, 開発用データに 9:1 でランダムで分割し, 残りの WMT-2017 のデータをテストデータとした. 言語対ごとのコーパスサイズを**表 1** に示す. コーパスに含まれる全ての言語から英語へ翻訳するデータ (all-en) のコーパスサイズと比較すると, de,ru,tr-en の言語ペアのみでは訓練データの数が $5360 \times 0.9 = 4824$ 文→ $2680 \times 0.9 = 2412$ 文と半分になっている.

また, XLM は XNLI[3] コーパスで事前学習されたモデル[*2]を使用した. MLP のハイパーパラメータは以下の組み合わせでグリッドサーチを行い, 最も開発データのピアソ

[*1] http://www.statmt.org/moses/index.php?n=Main.HomePage

[*2] https://github.com/facebookresearch/XLM

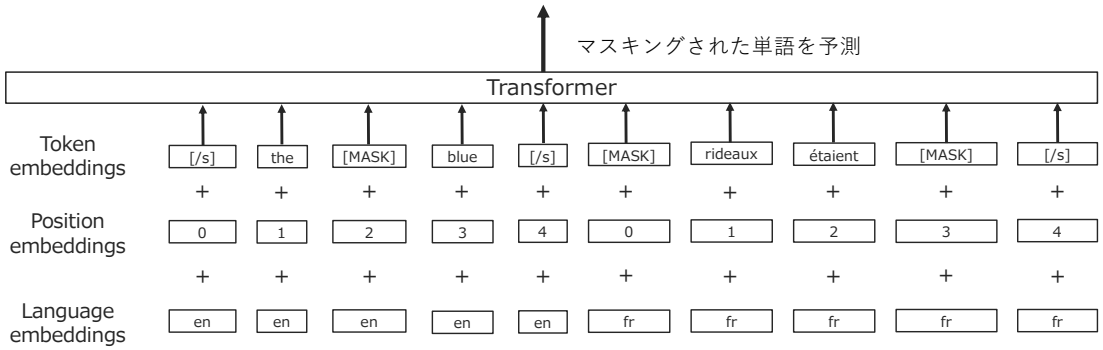
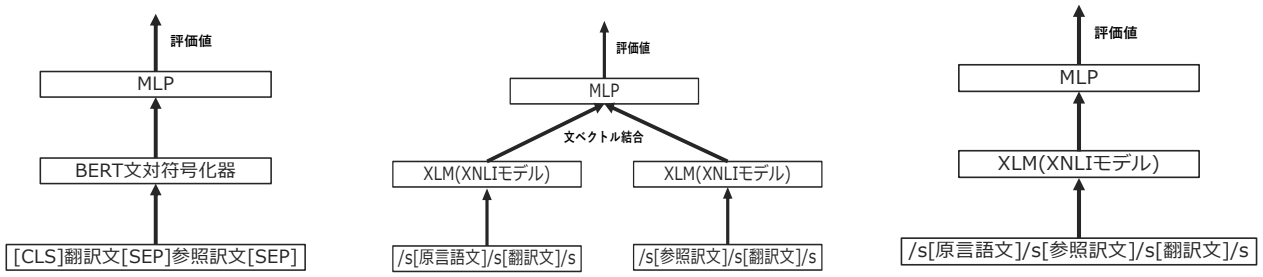


図 1 XLM のモデル図



(a) BERT-RUSE のモデル　(b) XLM による回帰モデル 提案手法 1　(c) XLM による回帰モデル 提案手法 2

図 2 XLM と BERT による回帰モデル図

ンの積率相関係数が高いモデルでテストを行った.

- バッチサイズ：8, 16
- 学習率・Adam：1e-4, 8e-5, 5e-5, 3e-5, 1e-5
- エポック数：1, ...,20
- 中間層の数：0
- ユニット数：2048
- ドロップアウト：0.1

評価実験の比較対象は, WMT Metrics Shared Task のベースラインである SentBLEU[2], WMT-2017 のデータで最高性能を記録した BERT ベースの RUSE[20] とし, それぞれの評価結果を引用した. また, 各自動評価手法のメタ評価はピアソンの積率相関係数に基づいて行った.

4.2 評価結果

実験結果を**表 2** に示す. XLM は de-en, ru-en, tr-en の三つの言語ペアにおいて BERT ベースの RUSE[20] が最も人手評価との相関性が高い結果となった. また提案手法 1 と提案手法 2 を比較すると, 三つの言語ペア全てにおいて提案手法 2 が上回っていることから, 原言語を XLM の入力とする際は, 原言語文-参照訳文-翻訳文と三文を同時に入力文とした方が MLP の学習が容易であることがわかる.

5. 分析

XLM を用いた提案手法が BERT ベースの評価手法に性能が及ばない原因を探るため, 提案手法 2 について学習曲線や, 評価値と人手評価値 (DA スコア [5]) の散布図, コーパスの DA スコア分布を出力し分析を行った.

de-en の言語ペアを例に学習曲線 (図 3, 4) を見てみると, 学習が進むともに誤差が小さくなりピアソンの相関係数が上昇していることがわかる. 訓練データのピアソンの相関係数が 7 epoch 目以降も上昇し続けているのに対して, 開発データやテストデータは頭打ちとなっている. ここで, 開発データとテストデータが頭打ちを迎えた後に誤差が大きくなったり, ピアソンの相関係数が小さくなったりと過学習の挙動を示していない. ru-en, tr-en 言語ペアについても同様に, 過学習をしているようには見られない**図 5**, **図 6**, **図 7**, **図 8**. これから, XLM を用いた回帰モデルはまだ改善の余地が残されている可能性がある.

また, ru-en の散布図 (図 10) を見ると, 0.0 よりも低い DA スコアにおいてモデルの評価値にばらつきがあることがわかる. 一方で, DA スコア分布図 (図 15- 17) では, 訓練データとテストデータでは DA スコアの分布に違いが見られ, テストデータは高いスコアに偏っている. ここで, de,ru,tr-en の DA スコア分布図 (図 21-23) を見るとスコアの偏りはないことから, 参照訳文と翻訳文よりも, モデルの評価値の低スコア帯におけるばらつきは, 原言語情報に影響を受けていると予想される. tr-en についても同様に散布図 11 は DA スコアがばらけており, tr-en の DA スコア分布図 (図 18-20) に影響を受けているのがわかる. これは, 今回の分析対象が提案手法 2(図 2(c)) で, 原言語文と目的言語文を同時に扱っているからだと考えられる. 提案手法 1 についても同様の分析を今後行いたい.

表 1　WMT-2017 Metrics Shared Task の to-English の人手評価付き文対数

	cs-en	de-en	fi-en	lv-en	ro-en	ru-en	tr-en	zh-en	de, ru, tr-en	all-en
WMT-2015	500	500	500	-	-	500	-	-	1000	2000
WMT-2016	560	560	560	-	560	560	560	-	1680	3360
WMT-2017	560	560	560	560	-	560	560	560	1680	3920
ALL	1620	1620	1620	560	560	1620	1120	560	4360	9280

表 2　WMT-2017 Metrics Shared Task 文単位のピアソンの相関係数値

	de-en	ru-en	tr-en	avg
SentBLEU[2]	0.432	0.484	0.538	0.484
BERT ベース RUSE[20]	**0.751**	**0.795**	**0.811**	**0.786**
提案手法 1	0.507	0.476	0.553	0.512
提案手法 2	0.565	0.618	0.619	0.600

6. おわりに

本研究では, 現在最も文レベルでの人手評価との相関性が高い自動評価手法である BERT ベースの RUSE[20] に原言語情報を擬似的に参照訳文として追加する手法を検討した. 提案手法は従来手法の性能に及ばなかったが, 今後更に分析を進め今後の改良につなげたい.

謝辞　本研究は JST さきがけ (JPMJPR1856) の支援を受けたものである.

参考文献

[1] Banerjee, S. and Lavie, A.: METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments, *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72 (2005).

[2] Bojar, O., Graham, Y. and Kamran, A.: Results of the WMT17 Metrics Shared Task, In Proceedings of WMT, pp. 489–513 (2017).

[3] Conneau, A., Rinott, R., Williams, G. L. A., Bowman, S. R., Schwenk, H. and Stoyanov, V.: Xnli: Evaluating cross-lingual sentence representations, In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing Association for Computational Linguistics, pp. 2475–2485 (2018).

[4] Doddington, G.: Automatic evaluation of machine translation quality using n-gram cooccurrence statistics, In Proceedings of the second international conference on Human Language Technology Research, p. 138145 (2002).

[5] Graham, Y., Mathur, N. and Baldwin, T.: Accurate Evaluation of Segment-level Machine Translation Metrics, In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics Human Language Technologies, pp. 1184–1191 (2015).

[6] Hiroki, S., Tomoyuki, K. and Mamoru, K.: RUSE: Regressor Using Sentence Embeddings for Automatic Machine Translation Evaluation, In Proceedings of WMT, pp. 764–771 (2018).

[7] Iyyer, M., Wieting, J., Gimpel, K. and Zettlemoyer, L.: Adversarial example generation with syntactically controlled paraphrase networks, In Proceedings of The 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, p. 18751885 (2018).

[8] Jacob Devlin, Ming-Wei Chang, K. L. and Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Vol. 1 (Long and Short Papers), pp. 4171–4186 (2018).

[9] Kusner, M. J., Sun, Y., Kolkin, N. I. and Weinberger, K. Q.: From Word Embeddings to Document Distances, *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, pp. 957–966 (2015).

[10] Lample, G. and Conneau, A.: Crosslingual Language Model Pretraining, p. arXiv:1901.07291 (2019).

[11] Ma, Q., Bojar, O. and Graham, Y.: Results of the WMT18 Metrics Shared Task: Both characters and embeddings achieve good performance, In Proceedings of WMT, p. 682701 (2018).

[12] Markus, D. and Marcu, D.: Hyter:Meaning-equivalent semantics for translation evaluation, the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, p. 162171 (2012).

[13] Papineni, K., Roukos, S., Ward, T. and Zhu, W.: Bleu: a method for automatic evaluation of machine translation, In Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, p. 311318 (2002).

[14] Qin, Y. and Specia, L.: Truly Exploring Multiple References for Machine Translation Evaluation, In Proceedings of the 18th Annual Conference of the European Association for Machine Translation (EAMT-15), p. 113120 (2015).

[15] Roy, A. and Grangier, D.: Unsupervised Paraphrasing without Translation, p. arXiv:1905.12752 (2019).

[16] 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)*, pp. 1715–1725 (2016).

[17] Snover, M., Dorr, B., Schwartz, R., Micciulla, L. and Makhoul, J.: A study of translation edit rate with targeted human annotation, In Proceedings of the Association for Machine Translation in the Americas, p. 223231 (2006).

[18] Tättar, A. and Fishel, M.: bleu2vec: the Painfully Familiar Metric on Continuous Vector Space Steroids, *Proceedings of the Second Conference on Machine Translation*, pp. 619–622 (2017).

[19] Vaswan, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. and Polosukhin, I.: Attention is all you need, In Advances in Neural Information Processing Systems, p. 60006010 (2017).

[20] 嶋中宏希，梶原智之，小町　守：BERT を用いた機械翻訳の自動評価，言語処理学会第 25 回年次大会 発表論文集，pp. 591–593 (2019).

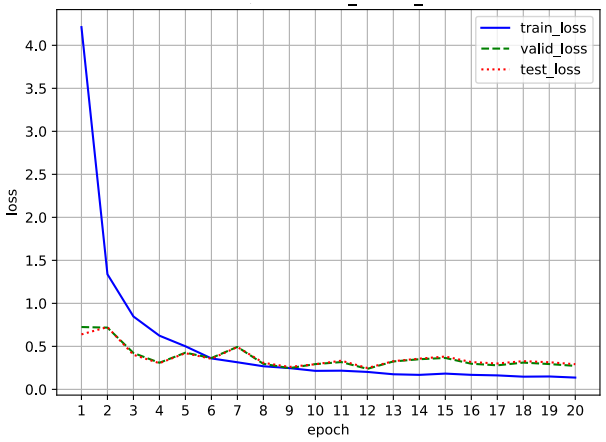


図 3　de-en の MSE 学習曲線

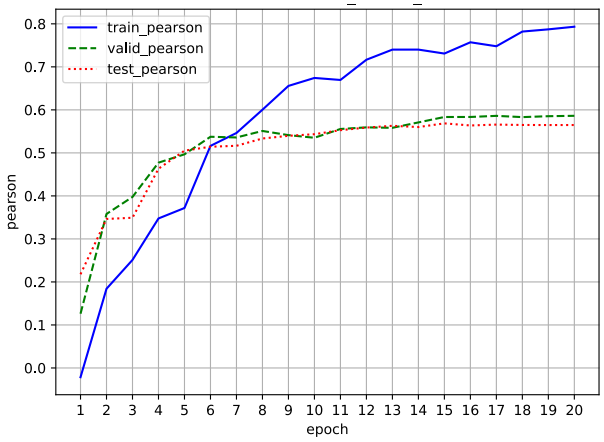


図 4　de-en のピアソン学習曲線

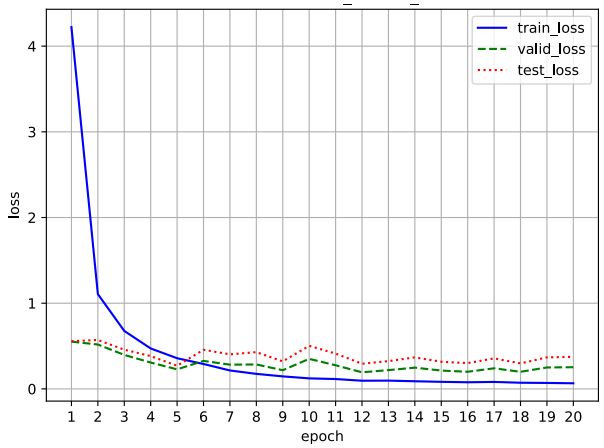


図 5　ru-en の MSE 学習曲線

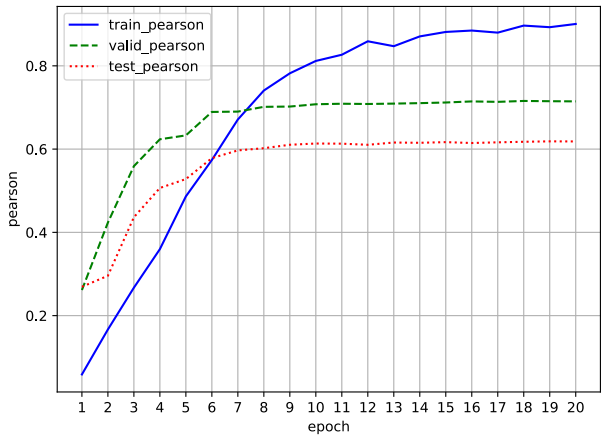


図 6　ru-en のピアソン学習曲線

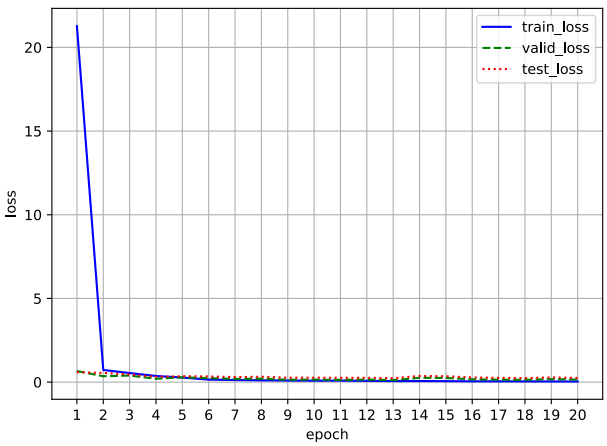


図 7　tr-en の MSE 学習曲線

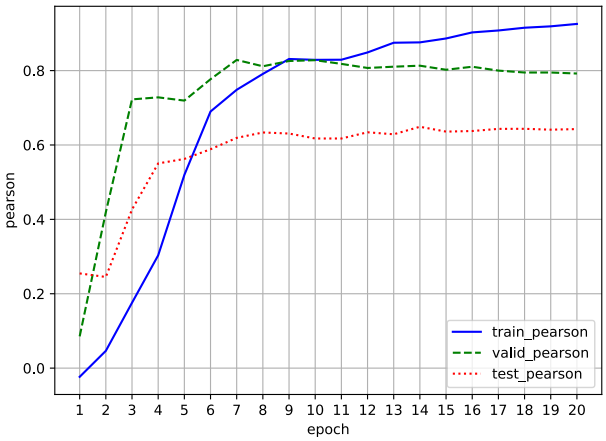


図 8　tr-en のピアソン学習曲線

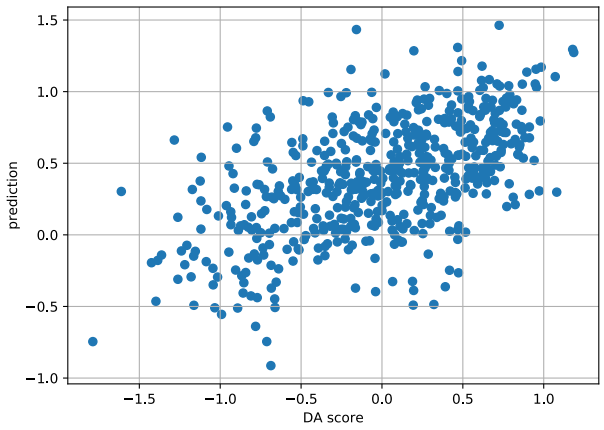


図 9　de-en のテスト時の DA[5] スコアと評価値の散布図

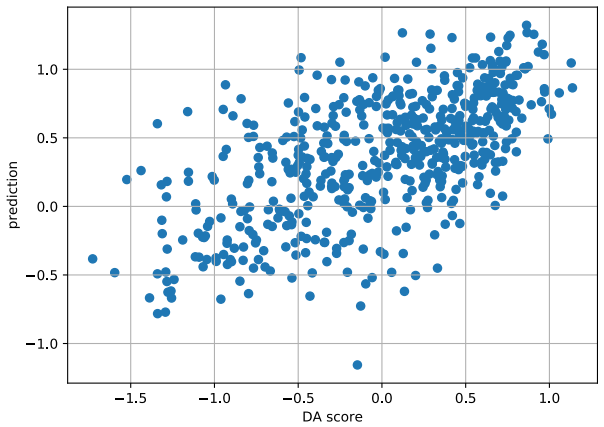


図 10　ru-en のテスト時の DA[5] スコアと評価値の散布図

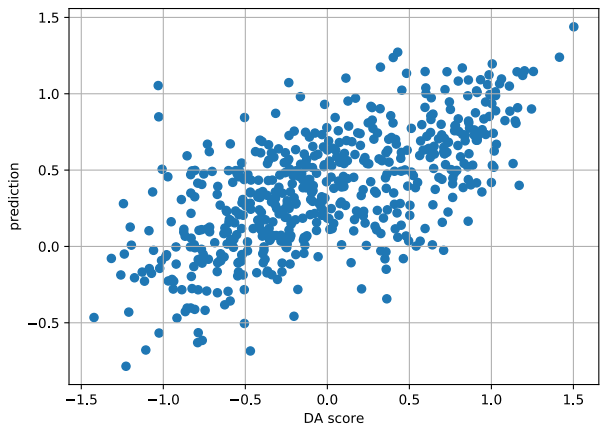


図 11　tr-en のテスト時の DA[5] スコアと評価値の散布図

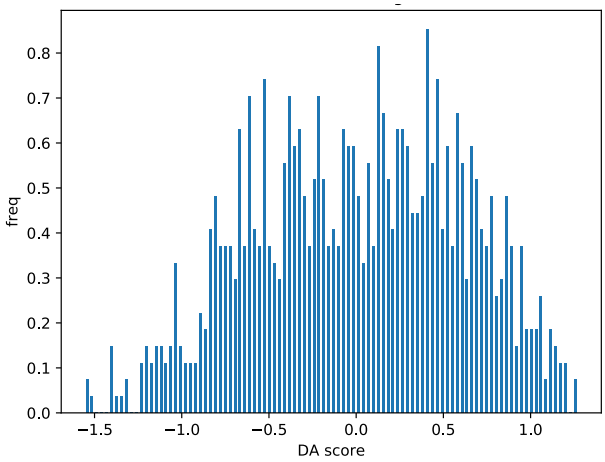


図 12　de-en の訓練データにおける DA[5] スコア分布

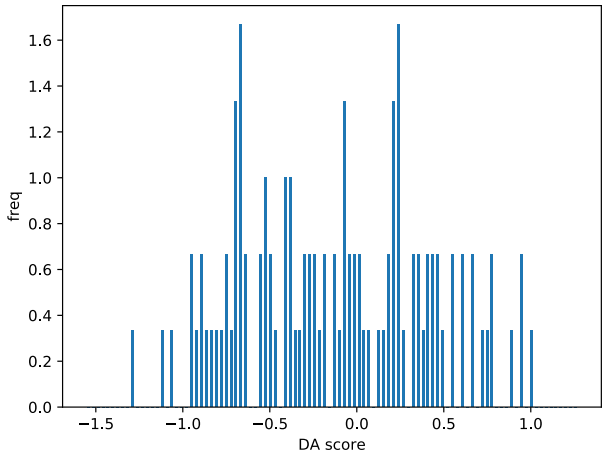


図 13　de-en の開発データにおける DA[5] スコア分布

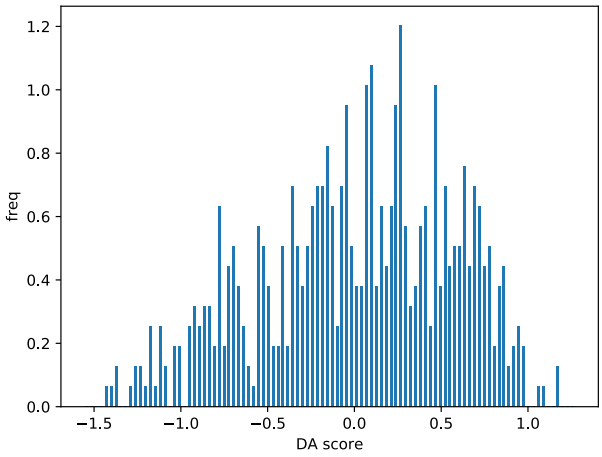


図 14　de-en のテストデータにおける DA[5] スコア分布

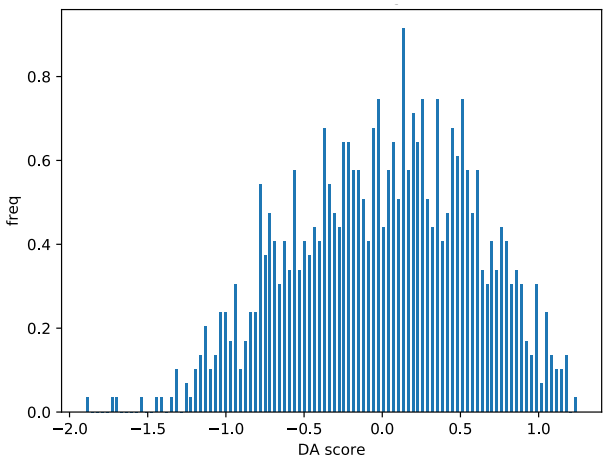


図 15　ru-en の訓練データにおける DA[5] スコア分布

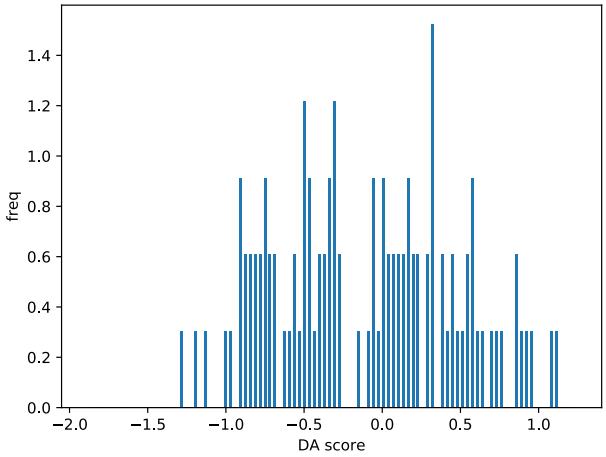


図 16　ru-en の開発データにおける DA[5] スコア分布

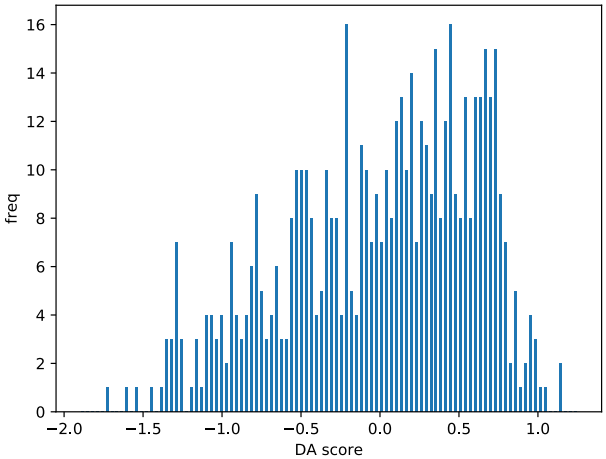


図 17　ru-en のテストデータにおける DA[5] スコア分布

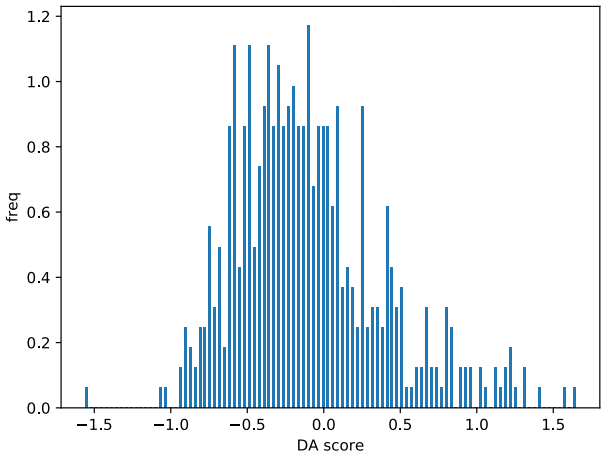


図 18　tr-en の訓練データにおける DA[5] スコア分布

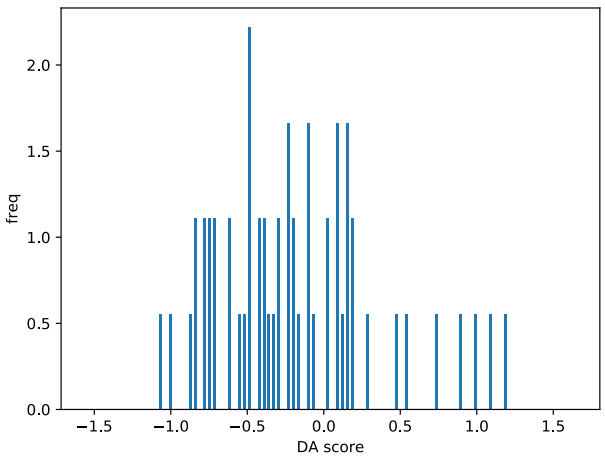


図 19　tr-en の開発データにおける DA[5] スコア分布

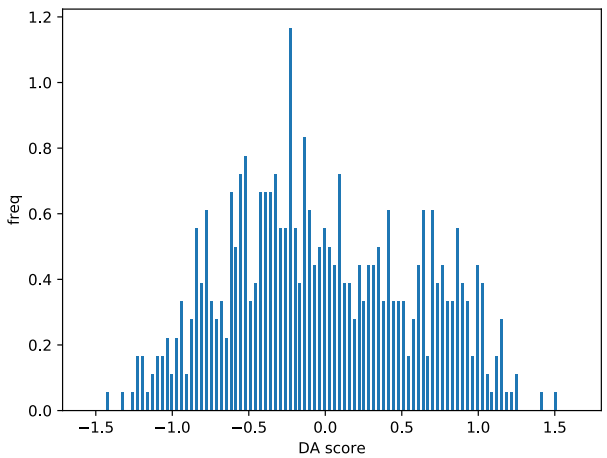


図 20　tr-en のテストデータにおける DA[5] スコア分布

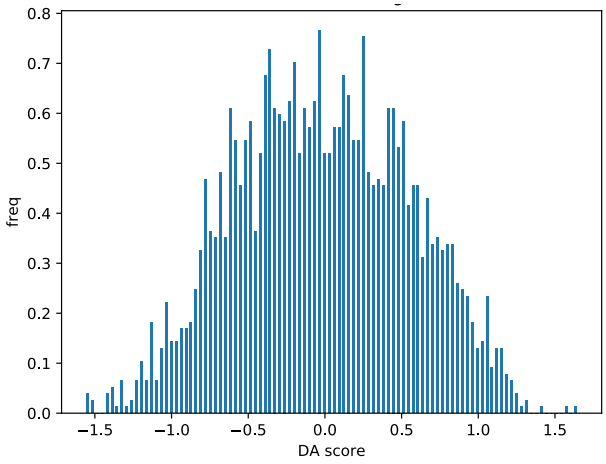


図 21 de,ru,tr-en の訓練データにおける DA[5] スコア分布

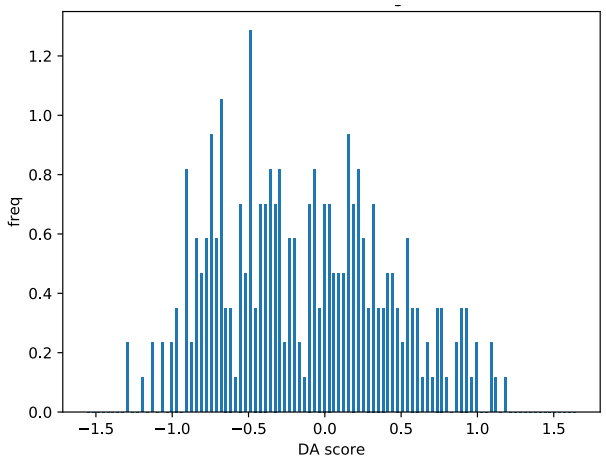


図 22 de,ru,tr-en の開発データにおける DA[5] スコア分布

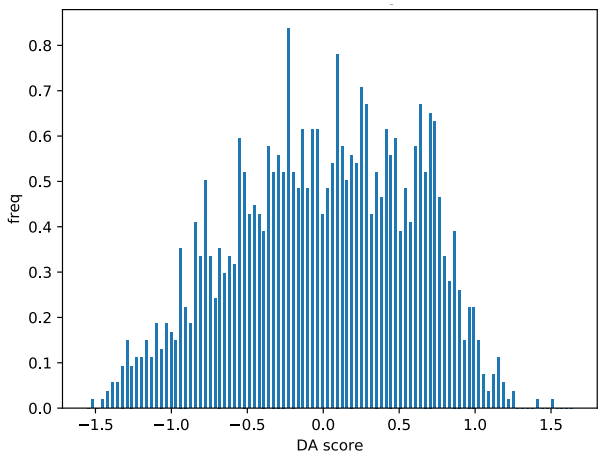


図 23 de,ru,tr-en のテストデータにおける DA[5] スコア分布