異なる難易度の参照文を用いる多段階難易度制御翻訳

谷 和樹 [1]　湯浅 亮也 [1]　田村 晃裕 [1]　梶原 智之 [2]　二宮 崇 [2]　加藤 恒夫 [1]
[1] 同志社大学　[2] 愛媛大学
{ctwh0176@mail4,ctwh0190@mail4,aktamura@mail,tsukato@mail}.doshisha.ac.jp
{kajiwara,ninomiya}@cs.ehime-u.ac.jp

概要

本研究では，目的言語文の難易度を多段階で制御する機械翻訳（多段階難易度制御翻訳）のための学習方法として，異なる難易度の参照文を用いる手法を提案する．従来手法では，原言語文と難易度付き目的言語文の文対の単位でモデルの学習を行うため，同一の原言語文に対応付く異なる難易度の目的言語文間を対比させた学習を行えない．そこで本研究では，学習対象の参照文と共に異なる難易度の参照文も使い，出力が学習対象の難易度から離れるほど大きな損失を与えるペナルティ項を導入した損失関数を用いる学習手法を提案する．日英多段階難易度制御翻訳の実験を行い，提案手法により BLEU が 0.66 ポイント改善できることを確認した．

1　はじめに

近年，ニューラル機械翻訳（NMT）はますます発展・普及し，利用者層が幅広くなっている．従来の一般的な NMT は，利用者や状況に依らない一律な翻訳を行う．一方で，近年では，出力される目的言語文の表現を制御するための研究が盛んになっている [1, 2, 3]．そのひとつに，ユーザの読解レベルにあわせた翻訳を行うため，原言語文と共に難易度を入力として受け付け，指定された難易度の目的言語文を生成する難易度制御翻訳がある．

初期の難易度制御翻訳では，難易度は 2 段階 [4] であったが，近年では，より柔軟に出力文の難易度を制御するため，3 段階以上の難易度（例えば，小学生，高校生，一般，専門家向けなど）を制御可能な多段階難易度制御翻訳（Multi-Level Complexity Controlling Machine Translation: MLCCMT）[5, 6] の研究が行われている．先行研究 [5] は英語-スペイン語間の MLCCMT に取り組み，マルチタスクモデルを提案している．マルチタスクモデルは，MLCCMT をメインタスクとし，難易度を制御しない通常の機械翻訳と同一言語内での平易化（単言語平易化）の 2 つのサブタスクと共にマルチタスク学習で学習されるモデルである．また，先行研究 [6] は日英 MLCCMT に取り組み，先行研究 [5] の手法を日英の言語対に適用している．

MLCCMT の学習では，表 1 のような，一つの原言語文が難易度の異なる複数の参照文に対応する教師データを使用できる．しかし，従来の学習では，原言語文と難易度付き目的言語文の文対の単位で学習を行う．例えば，表 1 の教師データは従来手法では 3 つの教師データ（日本語文-難易度 12 の英語文，日本語文-難易度 7 の英語文，日本語文-難易度 4 の英語文）に分解される．そして，教師データの各文対は独立に扱われる．そのため，難易度 12 の英語文への翻訳を学習する際，難易度 4 や難易度 7 の英語文と対比させた学習を行うことはできない．

そこで本研究では，同一の原言語文と難易度の異なる複数の参照文からなる組の単位で MLCCMT モデルを学習する手法を提案する．提案手法では，学習対象の参照文と共に異なる難易度の参照文も使い，出力が学習対象の難易度から離れるほど大きな損失を与えるペナルティ項を導入した損失関数に基づき MLCCMT モデルを学習する．これにより，例えば表 1 の教師データを用いて難易度 12 の英語文への翻訳を学習する際，出力を難易度 12 の英語文に最も近づけ，かつ，難易度 4 の英語文より難易度 7 の英語文に近づけるように学習を行う．

提案手法の有効性を先行研究 [6] で作成された評価データセットを用いた日英多段階難易度制御翻訳の実験で検証した．その結果，提案損失関数を利用することで BLEU が 0.66 ポイント改善でき，提案手法の有効性を確認した．

2　従来研究

本節では，MLCCMT の従来研究を述べる．先行研究 [5] では，MLCCMT の研究が初めて行われ，英

表 1　提案手法で用いる教師データの例

日本語文（最高難易度の英語文の Google 翻訳結果）	難易度	英語文（Newsela-auto コーパスのデータ）
科学者たちは、北極の波が氷氷をそのように壊す可能性があると想像していませんでした。	12	Scientists had never imagined that Arctic waves could break up pack ice so quickly.
	7	Scientists had never imagined that Arctic waves could break up ice so quickly.
	4	Scientists had never imagined that waves could break up ice so quickly.

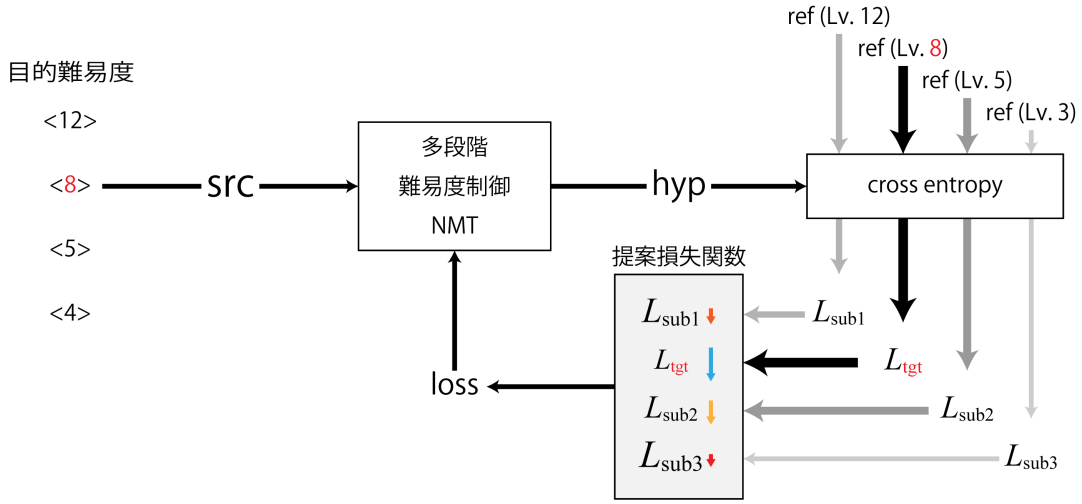

図 1　提案手法の概要図

語-スペイン語間の MLCCMT モデルとしてマルチタスクモデルが提案された．マルチタスクモデルでは，以下の式 (1) の損失関数に基づき，通常の機械翻訳と単言語平易化の 2 つのサブタスクとメインタスクの MLCCMT を 1 つのモデルで同時に学習する．

$$loss = L_{MT} + L_{Simplify} + L_{CMT} \tag{1}$$

$$L_{MT} = \sum_{(s_i,s_o)\in D_{MT}} CrossEntropy((s_i;\theta), s_o) \tag{2}$$

$$L_{Simplify} = \sum_{(s_o,c_{o'},s_{o'})\in D_S} CrossEntropy((s_o,c_{o'};\theta), s_{o'}) \tag{3}$$

$$L_{CMT} = \sum_{(s_i,c_o,s_o)\in D_{CMT}} CrossEntropy((s_i,c_o;\theta), s_o) \tag{4}$$

ここで，θ はモデルパラメータ，s_i は原言語文，s_o は目的言語文，$s_{o'}$ は s_o を平易化した文，$c_{o/o'}$ は $s_{o/o'}$ の難易度である．また，D_{MT}，D_S，D_{CMT} は，それぞれ，通常の機械翻訳，単言語平易化，MLCCMT 用の教師データである．

先行研究 [5] では，ベースラインの MLCCMT として，難易度を制御しない通常の機械翻訳モデルと単言語平易化モデルをパイプラインでつなげたパイプラインモデルが実装され，実験でマルチタスクモデルと比較された．その結果，マルチタスクモデルの方が性能が高いことが示されている．これらのモデルの教師データや評価データは，様々な難易度で記述された英語とスペイン語のニュース記事からなる Newsela コーパス[1] から自動作成されている．

先行研究 [6] は日英 MLCCMT に取り組んでいる．人手翻訳により日英 MLCCMT 用の評価データセットを作成し，先行研究 [5] のモデルを日英 MLCCMT に適用し，その性能を報告している．

これらの先行研究で使われている従来の MLCCMT は，原言語文と難易度付き目的言語文の文対の単位で学習されている．そのため，同一の原言語文に対応付く異なる難易度の目的言語文間を対比させた学習を行うことができない．

3　提案手法

本節では，学習対象の難易度の参照文と共に異なる難易度の参照文も用いて MLCCMT モデルを学習する手法を提案する．具体的には，指定された難易度の参照文に対する従来の損失と共に，出力と指定以外の難易度の参照文との損失が指定難易度の参照文との損失よりも小さくなることに対して，難易度が離れるほど大きなペナルティを与える損失関数に基づき MLCCMT モデルを学習する．提案手法の概要を図 1 に示す．

提案手法では，同一の原言語文と難易度の異なる複数の参照文からなる組の単位で MLCCMT モデルを学習する．表 1 に提案手法で用いる教師データの例を示す．実験では，英語の多段階難易度コーパス

1）　https://Newsela.com/data/

Newsela-auto [7] の最高難易度の英語文を Google 翻訳で日本語文に翻訳することで日英 MLCCMT の教師データを作成した．作成した教師データは，一つの日本語文に対して難易度がそれぞれ異なる 3 から 5 文の英語文が対応付けられている．提案手法では，そのまとまり（組）単位で学習を行う．

具体的には，以下の式 (5) の損失関数を最小化することでモデルを学習する．提案損失関数は，同一の原言語文に対応付けられた難易度の異なる複数の参照文をまとめた組単位で定義される．

$$loss = L_{\mathrm{tgt}} + \alpha \cdot \frac{1}{n-1} \sum_{k=1}^{n-1} {d_k}^2 \cdot \max(L_{\mathrm{tgt}} - L_{\mathrm{sub}_k}, 0) \quad (5)$$

$$L_{\mathrm{tgt}} = CrossEntropy((x, c_{\mathrm{tgt}}; \theta), y_{\mathrm{tgt}}) \quad (6)$$

$$L_{\mathrm{sub}_k} = CrossEntropy((x, c_{\mathrm{tgt}}; \theta), y_{\mathrm{sub}_k}) \quad (7)$$

ここで, tgt と sub_k は, それぞれ, 学習対象の参照文と同一組内の学習対象以外の参照文を表す．n を組内の参照文数とすると, sub_k の数は $n-1\,(k = 1, ..., n-1)$ である．そして，$L_{\mathrm{tgt}}, L_{\mathrm{sub}_k}, x, y, c, \theta, \alpha, d_k$ は，それぞれ，学習対象の難易度の参照文との誤差，学習対象以外の難易度の参照文との誤差，原言語文，目的言語文，難易度，モデルパラメータ，ハイパーパラメータ，難易度の差（$d_k = c_{\mathrm{tgt}} - c_{\mathrm{sub}_k}$）である．

提案損失関数は，第 1 項目の L_{tgt} により，出力を学習対象の参照文に近づける．この項は，従来の学習で用いられている損失と同じである．これに加えて第 2 項目で，出力が組中の参照文の中で学習対象の参照文に最も近づくように，$L_{\mathrm{tgt}} - L_{\mathrm{sub}_k} > 0$ の時にペナルティを与える．この際，表 1 の例のように，難易度が離れるほど語彙や構文が変化して出力が異なる傾向があるため，難易度の差の二乗（${d_i}^2$）を掛けて，学習対象の難易度から離れた難易度の参照文ほど大きなペナルティを与える．また，学習対象以外の参照文数（$n-1$）は組によって異なるため，その数で割る．このペナルティ項の影響度はハイパーパラメータ α で制御する．$\alpha = 0$ のときは学習対象以外の参照文を用いない従来の学習手法と同等になる．

4 実験

4.1 実験設定

本実験では，日英多段階難易度制御翻訳の実験により，提案手法の有効性を検証する．評価データは先行研究 [6] で作成された日英 MLCCMT 用評価データ（1,014 組）を用いた．教師データは D_{MT}, D_{CMT}, D_S の 3 種類のデータを用いた．D_{MT} は，JParaCrawl [8] と News-Commentary の日英対訳文対からなる機械翻訳用データであり，D_S は，Newsela-auto [7] 中の英語文集合毎に抽出した最高難易度の英語文とそれ以外の英語文の対からなる単言語平易化用データである．そして，D_{CMT} は，Newsela-auto [7] 中の各英語文集合に対して，最高難易度の英語文を Google 英日翻訳で翻訳した日本語文を付与した MLCCMT 用データ（ただし，評価データセット [6] に含まれるデータは除く）である．各教師データのデータ量は付録 A の表 4 に示す．

本実験では提案手法の有効性を検証するため，3 節で説明した提案手法と，提案手法においてハイパーパラメータ α を 0 にしたベースライン手法の性能を比較する．また，2 節で述べた従来のマルチタスクモデルの性能とも比較する．各モデルは Transformer モデル [9] を採用した．実装は Fairseq [10] を用いた．

マルチタスクモデルでは，先行研究 [6] と同様に，日英 MLCCMT をメインタスク，通常の日英翻訳と英語多段階平易化をサブタスクとしたマルチタスク学習を行った．日英 MLCCMT タスクの学習には D_{CMT}，日英翻訳タスクと英語多段階平易化タスクの学習には，それぞれ，D_{MT} と D_S を使用した．その他の実験設定は付録 A の表 5 に示す．

提案手法では，まず，D_{MT} を用いて日英 NMT モデルの事前学習を行った．この事前学習の設定は，Kiyono ら [11] に倣った．その後，D_{CMT} を用いて 3 節の提案損失関数により，事前学習したモデルをファインチューニングした．ハイパーパラメータ α のチューニングは，検証データに対する BLEU [12] の値に基づいて行った．その結果，提案手法の α は 0.5 とした．チューニングの詳細は付録 A の表 6 に示す．その他の実験設定は付録 A の表 5 に示す．

評価指標は，先行研究 [6] 同様，翻訳性能の指標として BLEU [12]，平易化の指標として SARI [13] を用いた．また，目的難易度の出力文集合から算出した fkgl [14] と目的難易度との平均絶対誤差である MAE_{fkgl} [15] も用いた．これらの指標の数値は，EASSE [16] を用いて算出した．

4.2 実験結果

実験結果を表 3 に示す．提案手法及びマルチタスクモデルの性能は，ランダムシードを変更して学習

表 2 翻訳結果の比較（BS：ベースライン手法（$\alpha = 0.0$），従来：マルチタスクモデル，提案：提案手法）

難易度		
12	src	チェンバース氏は、敵対する立場の議員が話すときはペンを置き、質問を浴びせる。
	ref	When his adversaries speak, Chambers puts down the pen and peppers them with questions.
	BS	Chambers said he put pens and posed questions when lawmakers in opposition positions talk.
	従来	Chambers said he puts a pen when lawmakers in hostile positions speak, and poses questions.
	提案	Chambers said he puts a pen on when a hostile senator speaks and puts a question out.
7	ref	When people he opposes or doesn't agree with speak, Chambers puts down the pencil and instead throws questions at them.
	BS	Chambers said opposition lawmakers put pens when they talk and pose questions.
	従来	Chambers said he puts a pen when lawmakers in hostile positions speak, and takes questions.
	提案	Chambers put a pen on when opposing lawmakers talk and take up questions.
6	ref	When people are speaking who Chambers does not support, he stops drawing and then asks the presenter questions.
	BS	Chambers said opposition lawmakers put pens when they talk and pose questions.
	従来	Chambers said he puts a pen when lawmakers in hostile positions speak.
	提案	Chambers put a pen on when opposing lawmakers talk and take up questions.
4	ref	If he doesn't agree with something someone says, he will stop drawing and ask questions instead.
	BS	Chambers said he put pens on when lawmakers talk against him.
	従来	Chambers said he puts a pen when lawmakers speak against him.
	提案	Chambers puts pens on when lawmakers talk, he said.
3	ref	If he doesn't agree with something he hears, he will stop drawing and ask questions instead.
	BS	They put pens on when they talk.
	従来	She puts a pen on the floor.
	提案	Chambers puts pens on when lawmakers talk.

表 3 実験結果

モデル	BLEU(%) ↑	SARI(%) ↑	MAE_{fkgl} ↓
パイプライン [6]	15.12	23.89	2.084
マルチタスク [6]	20.17	26.78	0.600
マルチタスク	22.80	28.06	0.838
ベースライン ($\alpha = 0$)	21.97	27.70	0.768
提案手法 ($\alpha = 0.5$)	22.64	28.05	0.793

した 3 回の実験結果の平均の値である．表 3 より，提案手法の方がベースライン手法（$\alpha = 0.0$）の性能よりも高いことが分かる．これより，提案損失関数に基づき，学習対象以外の参照文も用いて学習することで MLCCMT の性能を改善できることが分かり，提案手法の有効性を確認できた．

また表 3 より，提案手法は従来研究 [6] のパイプラインモデルやマルチタスクモデルよりも高い性能を達成できたことが分かる．ただし，従来研究のマルチタスクモデル [6] の教師データ量は本実験で使用した教師データ量よりも少ないことに注意されたい（付録 A の表 4 参照）．提案手法と教師データ量を揃えた本実験のマルチタスクモデルには，提案手法はわずかに及ばない結果となった．

5 考察

本節では，各モデルが出力した翻訳結果の例を比較する．例としてテストデータの中で最も難易度の段階数が多い 5 段階の組から抽出した翻訳例を表 2 に示す．表 2 において，BS がベースライン手法（$\alpha = 0$），従来がマルチタスクモデル，提案が提案手法である．マルチタスクモデルと提案手法では，異なるランダムシードを用いて 3 回実験を行ったため，3 つの翻訳結果が得られる．表 2 では，BLEU の値が最も良かったシードを用いた場合の翻訳結果を示している．

表 2 より，ベースライン手法は目的の難易度が異なる場合にも同じ出力をする場合がある．その一方で提案手法は，従来手法のうち最高性能であるマルチタスクモデルと同様に，目的の難易度ごとに適切に異なる出力を行っていることが分かる．これより，提案損失関数によって異なる難易度の参照文を対比させて学習させることで，難易度を制御しやすくなることが実例により確認できた．

6 まとめ

本研究では，多段階難易度制御翻訳のための学習手法として，学習対象の難易度の参照文と共に異なる難易度の参照文も用いて，出力が学習対象の難易度の参照文から離れるほど大きな損失を与えるペナルティ項を導入した損失関数に基づき学習する手法を提案した．そして，日英多段階難易度制御翻訳の実験を通じて提案手法の有効性を確認した．今後は，マルチタスクモデルと提案損失関数を組み合わせることで更なる性能改善が実現できるかを検証する予定である．

謝辞

本研究は JSPS 科研費 JP22K12177，JP21K12031 の助成を受けたものである．また，本研究成果の一部

は，国立研究開発法人情報通信研究機構の委託研究 (No. 225) により得られたものである．ここに謝意を表する．

参考文献

[1] Rico Sennrich, Barry Haddow, and Alexandra Birch. Controlling politeness in neural machine translation via side constraints. In **Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 35–40, 2016.

[2] James Kuczmarski and Melvin Johnson. Gender-aware natural language translation. In **Technical Disclosure Commons, (October 08, 2018)**, 2018.

[3] Andrea Schioppa, David Vilar, Artem Sokolov, and Katja Filippova. Controlling machine translation for multiple attributes with additive interventions. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 6676–6696, 2021.

[4] Kelly Marchisio, Jialiang Guo, Cheng-I Lai, and Philipp Koehn. Controlling the reading level of machine translation output. In **Proceedings of Machine Translation Summit XVII: Research Track**, pp. 193–203, 2019.

[5] Sweta Agrawal and Marine Carpuat. Controlling text complexity in neural machine translation. 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. 1549–1564, 2019.

[6] Kazuki Tani, Ryoya Yuasa, Kazuki Takikawa, Akihiro Tamura, Tomoyuki Kajiwara, Takashi Ninomiya, and Tsuneo Kato. A benchmark dataset for multi-level complexity-controllable machine translation. In **Proceedings of the Thirteenth Language Resources and Evaluation Conference**, pp. 6744–6752, Marseille, France, June 2022. European Language Resources Association.

[7] Chao Jiang, Mounica Maddela, Wuwei Lan, Yang Zhong, and Wei Xu. Neural CRF model for sentence alignment in text simplification. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 7943–7960, 2020.

[8] Makoto Morishita, Jun Suzuki, and Masaaki Nagata. JParaCrawl: A large scale web-based English-Japanese parallel corpus. In **Proceedings of the 12th Language Resources and Evaluation Conference**, pp. 3603–3609, Marseille, France, May 2020. European Language Resources Association.

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, L ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In **Advances in Neural Information Processing Systems**, Vol. 30, 2017.

[10] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)**, pp. 48–53, 2019.

[11] Shun Kiyono, Takumi Ito, Ryuto Konno, Makoto Morishita, and Jun Suzuki. Tohoku-AIP-NTT at WMT 2020 news translation task. In **Proceedings of the Fifth Conference on Machine Translation**, pp. 145–155, 2020.

[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, 2002.

[13] Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. Optimizing statistical machine translation for text simplification. **Transactions of the Association for Computational Linguistics**, Vol. 4, pp. 401–415, 2016.

[14] J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch, 1975.

[15] Daiki Nishihara, Tomoyuki Kajiwara, and Yuki Arase. Controllable text simplification with lexical constraint loss. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop**, pp. 260–266, 2019.

[16] Fernando Alva-Manchego, Louis Martin, Carolina Scarton, and Lucia Specia. EASSE: Easier automatic sentence simplification evaluation. 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): System Demonstrations**, pp. 49–54, Hong Kong, China, November 2019. Association for Computational Linguistics.

A 付録

本付録では，教師データの統計量を表 4 に示し，実験で用いたモデルの詳細設定を表 5 に示す．そして，提案手法のハイパーパラメータ α のチューニング結果を表 6 に示し，提案手法とマルチタスクモデルのランダムシードごとの実験結果を，それぞれ，表 7 と表 8 に示す．

表 4 各モデルで使用する教師データ（文対数）

モデル	D_{MT}	D_S	D_{CMT}
パイプライン [6]	9.7M	150K	
マルチタスク [6]	3M	200K	200K
マルチタスク	9.7M	260K	260K
ベースライン及び提案手法	9.7M		260K

表 5 Fairseq を用いた実験設定

	マルチタスク	ベースライン及び提案手法	
architecture	Transformer [9]	Transformer [9]	
optimizer	adam	adam	
adam-betas	(0.9, 0.98)	(0.9, 0.98)	
dropout	0.1	0.0	
batch-size	50	200	
patience	10	10	
max-epoch	100	100	
bpe	sentencepiece	sentencepiece	
language	en + ja	en	ja
character coverage	0.9995	1.0	0.9998
vocab size	32,000	32,000	32,000

表 6 ハイパーパラメータ α のチューニング（検証データにおける性能）

α	0.0	0.5	1.0	2.5	5.0	7.5	10.0	12.5
BLEU	28.11	**28.31**	28.09	27.41	27.73	27.92	28.08	27.88

表 7 提案手法 ($\alpha = 0.5$) のランダムシード毎の実験結果

シード	BLEU(%) ↑	SARI(%) ↑	MAE_{fkgl} ↓
1	22.38	27.90	0.69
2	22.58	28.07	0.84
3	22.94	28.18	0.86

表 8 マルチタスクモデルのランダムシード毎の実験結果

シード	BLEU(%) ↑	SARI(%) ↑	MAE_{fkgl} ↓
1	23.04	28.16	0.77
2	23.04	28.15	0.93
3	22.33	27.89	0.82