深刻な誤訳の識別に向けた分類型翻訳評価データセットの構築

須藤克仁 [1,2]　髙橋洸丞 [1]　中村哲 [1]
[1] 奈良先端科学技術大学院大学　[2] 科学技術振興機構さきがけ
{sudoh, takahashi.kosuke.th0, s-nakamura}@is.naist.jp

1 はじめに

ニューラル機械翻訳 (NMT) 技術の進展によって，機械翻訳の平均的な訳質は従前と比較して大きく向上した．NMT の訳は流暢性に優れることが特徴であるが，その一方で重要語句の訳抜けや誤訳が問題視されている．これはかつての統計的機械翻訳 (SMT) による「見るからに機械翻訳」という読みづらくぎこちない文ではなく，一見して問題がないような文が機械翻訳の結果として得られるようになったが故に，翻訳の評価において内容の過不足や文意の正確性により注意を払わねばならないことの現れともいえる．

機械翻訳評価の事実上の標準である BLEU[1] は n-gram 精度と簡潔ペナルティによって訳語の表層的な正確性とカバレージを粗く見積もる指標であり，訳文の文意の正確性の評価にはほど遠い．他方，分散表現を用いることで表層一致によらない翻訳評価の手法として BERT Regressor [2], BERTScore [3], MoverScore [4], BLEURT [5] 等が提案され，人手評価と高い相関があることが示されている．現在の翻訳人手評価の主流は直接評価 (Direct Assessment; DA) [6] と呼ばれる標準化された単一の人手評価値である．DA の予測を回帰として解くアプローチは上記 BERT Regressor や BLEURT でも用いられており，人手評価を自動評価で再現するという観点では理にかなった方法である．

しかしながら，DA は 1 次元の評価指標であって，様々な誤訳を含む翻訳結果を評価するには不十分である．例えば，「私は 1 万円支払った」という参照訳に対して，(1)「私は 1 万円支払わなかった」，(2)「支払い円 1 万した」，(3)「彼は財布を拾った」はいずれも誤訳であることは確かだが，DA の枠組みでどう異なる評価値を付するかの判断は容易でない．また，(1) は一見して多くの訳語は正しいものの文意としては大きく異なり，実用上深刻な誤訳といえる．こうした誤訳の識別は自動評価にとっては難しい課題である．

本研究では，「訳文の文意が読み取れるか」「訳文の文意が参照訳と比較し誤解を招かないか」の二点に着目した分類型の翻訳評価を目指す．複数の視点による翻訳評価はかつて広く利用されていた 5 段階の流暢さ (Fluency) と忠実さ (Adequacy) に基づく人手評価 [7] や，NTCIR PatentMT の人手評価 [8] で採用された 5 段階の忠実さと受容度 (Acceptability)，多数の観点を用いた評価指標である多次元評価 (Multidimentional Quality Metrics; MQM) [9] 等がある．忠実さは文意の伝達量に基づいており，誤訳の影響を測るような定義となっていない．受容度は忠実さの評価が 5 であった訳文についてのみ文法的な正しさや流暢さを分類するものであり，深刻な誤訳を識別するには適さない．多数の観点は評価の詳細化のために有益であるが，人手評価が煩雑になりやすい．また，本研究と独立に，Popović は同様の方針により解釈性 (Comprehensibility) と忠実さ (Adequacy) に影響を与える語句を Major と Minor の二段階で分類する人手評価アノテーションを実施している [10, 11]．語句単位での評価アノテーションは誤り要因の同定に繋がり有益であるものの，忠実さの面で語句単位の誤りが文意に与える影響は明示的に考慮されていない．本研究では語句単位ではなく文単位での人手評価・自動評価を目的とし，文意解釈性，文意正確性[1)]に基づく文単位での分類型翻訳評価を提案する．提案する評価方法では，まず文意が読み取れるかを文意解釈性として評価し，その読み取った文意が正しいかどうかを参照文との比較によって文意正確性として評価する．

本稿では，上記の観点に基づく評価基準とそれに基づく翻訳評価データセットの構築について述べ，このデータセットを用いた自動評価実験の結果を示す．

2 分類型翻訳評価基準

提案する評価基準は流暢さ・忠実さと類似するが，それらがすべて段階的な設計となっていたのに対し，文意正確性については誤訳の分類のために段階的でない分類型の設計を行った．以下，詳細を述べる．

1) 過去の 5 段階評価に基づく「忠実さ」との混同を避けるため本稿では異なる用語を用いる．

表 1　文意解釈性の評価基準

ラベル	説明
F	何を伝達しようとしているかが理解できない箇所がある（言葉遣いとして内容伝達に失敗している場合を指し、専門用語の意味が分からない等は除く）
D	表記や文法の誤り、表現の問題でしっかり読まないと伝達内容が理解できない
B	表記や文法の誤りがあるが、伝達内容の理解は容易
A	文法的に正しいが、不自然な表現がある
S	文法的に正しく、言葉遣いも自然である

表 2　文意正確性の評価基準

ラベル	説明
N	文意が理解できず、正確性評価に値しない
O	参照文と無関係な内容が伝達されている
C	参照文と矛盾する内容が伝達されている
F	参照文と矛盾とまではいかないが重要な情報の誤りや過不足があり文意の重大な誤解が起こり得る
B	参照文と文意に若干の齟齬はあるが、大きな誤解を招くほどではない
A	参照文と文意に僅かな違いがあるが、ほぼ誤解の心配はない
S	参照文と文意が同一と考えて差し支えない

2.1　文意解釈性

表 1 に文意解釈性の評価基準を示す．従来の流暢さの基準 [7] では最低の 1 が理解不能 (Incomprehensible) である他は流暢性のみが着目され，ARPA における機械翻訳評価 [12] で検討された解釈性の観点は考慮されていなかった．本研究では流暢性と解釈性を一つの基準で表現するため，(1) 表記や文法，言語表現の問題が訳文の解釈を不可能にしている場合 (F)，(2) 解釈が不可能ではないが容易とは言い難い場合 (D)，(3) 解釈が容易である場合 (B, A, および S) の大きく 3 つに分け，(3) については流暢性や文法適格性の影響の違いで 3 段階，の 5 段階の基準とした．

2.2　文意正確性

表 2 に文意正確性の評価基準を示す．従来の忠実さの基準 [7] は意味内容がどの程度伝達されたか，という量的な尺度であったのに対して，本研究では誤訳の深刻さによる伝達内容の誤解リスクを評価基準の中心とし，(1) そもそも内容が解釈できない場合 (N)，(2) 誤訳により誤解を招き得る場合 (O, C, および F)，(3) 誤解の心配がない，もしくは軽微な場合 (B, A, および S) の大きく 3 つに分け，(2) については誤訳のタイプの違いで 3 種類，(3) については誤解を招く程度の違いで 3 段階，の計 7 種類の基準とした．

表 3　人手評価におけるアノテータ間一致度（Kappa 係数 (κ) およびラベル一致率 (r)）

Metric		A-B	A-C	B-C
文意解釈性	κ	0.2860	0.3773	0.2489
	r	0.4512	0.5113	0.4014
文意正確性	κ	0.3947	0.2684	0.2774
	r	0.5459	0.5870	0.5752

3　評価データセットの構築と分析

文意解釈性と文意正確性に基づき作成した機械翻訳の人手評価データセットについて詳細を述べる．

3.1　人手評価

本評価データセットは，WMT 2015 から WMT 2017 までの評価尺度 (Metrics) タスクのデータ（訳文・参照訳合計 9,280 文対，いずれも英語）を利用し，文意解釈性と文意正確性のアノテーションを人手で付したものである．アノテーションは英語への翻訳業務の経験を持つ 3 名のアノテータが独立に，英語の訳文と参照訳のみを見る形で行った．Popović は原文との比較評価を行ったが，参照訳の品質に影響を受けない一方で両言語に堪能な評価者を要し言語拡張性に劣る点が問題といえる．

3.2　評価データセットの分析

構築した評価データセットをアノテータ一致度と DA スコアとの関係の二点で分析した．

3.2.1　アノテータ間一致度

表 3 にアノテータ間一致度を Kappa 係数 (κ) とラベル一致率 (r) で示す．一致度は十分高いとは言い難いが，過去の WMT 人手評価データ [13, 14] と同等程度である．A-B や B-C で文意解釈性の一致度は文意正確性よりも低くなっているが，これはアノテータ B が A や C より文意解釈性の評価が厳しかったこと（詳細は付録の表 10-11 参照）によるものと考えられる．

3.2.2　文意解釈性・文意正確性と DA スコアの関係

人手評価のラベルと DA スコアとの関係を各ラベルに対応する翻訳結果の DA スコアの平均と標準偏差で

表 4　評価ラベルに対する平均 DA スコア（文意解釈性）

	A	B	C
F	-0.644±0.371	-0.692±0.356	-0.649±0.378
D	-0.421±0.408	-0.420±0.399	-0.400±0.418
B	-0.079±0.478	0.019±0.474	-0.129±0.449
A	0.165±0.479	0.408±0.485	0.122±0.467
S	0.428±0.524	0.644±0.465	0.427±0.521

表 5　評価ラベルに対する平均 DA スコア（文意正確性）

	A	B	C
N	-0.646±0.369	-0.692±0.356	-0.662±0.373
O	-0.990±0.367	-0.926±0.415	-0.963±0.363
C	-0.370±0.460	-0.366±0.468	-0.200±0.501
F	-0.453±0.438	-0.499±0.425	-0.279±0.473
B	-0.076±0.435	-0.092±0.417	-0.029±0.414
A	0.417±0.361	0.414±0.363	0.347±0.414
S	0.814±0.278	0.839±0.294	0.756±0.327

表したものを表 4, 5 に示す．各ラベルに対応する DA スコア範囲はアノテータ間で若干異なるものの，おおまかな順序はほぼ同じになっていることが分かる．特に順序尺度に近い設計となっている文意解釈性の各ラベルと文意正確性の B, A, S については DA スコアでもその優劣と一致する結果が得られている．

また，順序尺度となっていない文意正確性の N, O, C, F の各ラベルについては，無関係な内容を表す O が最も DA スコアが低く，その次に理解不能な訳文を表す N が低くなっており，C や F は誤解を招くリスクが高い深刻な誤訳でありながら DA スコアが O や N より平均的には高くなっている．DA スコアという一次元の尺度でこうした異なる種類の誤訳を識別することには限界があり，読めない誤訳，読めるが全く無関係な内容の誤訳，内容は一見類似しているが誤解を招くような誤訳，文意の伝達にはあまり影響を与えないような誤訳，を識別するためには，本研究のような複数の評価観点と分類型の評価の導入が重要である．

4　自動評価実験

本研究で提案する分類型機械翻訳評価を自動化できるかを検証するため，評価データセットをもとに以下の自動評価実験を行った．

4.1　データ

評価データセットのうち WMT 2017 のデータに相当する 3,920 文対（7 言語から英語への翻訳データが言語対ごとに 560 文対）をテストセットとして利用

表 6　自動評価実験での評価ラベル統計（文意解釈性）

	学習	開発	テスト
F	545	74	282
D	992	96	602
B	1,655	196	1,341
A	808	80	899
S	824	90	796

表 7　自動評価実験での評価ラベル統計（文意正確性）

	学習	開発	テスト
N	617	87	350
O	15	2	19
C	93	6	40
F	1,161	108	717
B	1,433	162	1,441
A	1,208	143	1,165
S	297	28	188

し，残りのデータからランダムに選んだ 536 文対を開発セット，その残りの 4,824 文対を学習セットとして利用した．評価データセットには 3 名の異なるアノテータによるラベルが付されているため，本実験では以下のヒューリスティクスに基づいて単一のラベルを選択して利用した．

- 2 名以上の付与したラベルが一致していればそれを用いる
- ラベルの一致がなければ，最も悪い評価ラベルを利用する．ラベルの順序は文意解釈性では F < C < B < A < S，文意正確性では C < F < N < O < B < A < S とする．

表 6, 7 に学習・開発・テストセットにおける評価ラベルの統計を示す．

4.2　実験設定

本研究の自動評価は機械翻訳結果と参照訳を入力とし評価ラベルを予測する分類問題として定式化される．本実験では予測モデルは文意解釈性と文意正確性で独立に学習した．

予測モデルの実装には HuggingFace Transformers[2] と学習済みの RoBERTa モデル (`roberta-large`) [15] を利用し，RoBERTa の `[CLS]` トークンに対する最終層のベクトルを入力とする全結合層 1 層を用いた分類モデルを構築し，交差エントロピーを損失関数として学習を行った．ここで，表 6, 7 でも明らかなようにラ

2)　https://github.com/huggingface/transformers

表 8　文意解釈性予測の混同行列．全体の正解率は 0.578.

正解\予測	F	D	B	A	S
F	**206**	45	22	8	1
D	45	**266**	250	43	4
B	15	134	**782**	358	52
A	2	11	187	**560**	139
S	0	2	35	306	**453**

表 9　文意正確性予測の混同行列．全体の正解率は 0.600.

正解\予測	N	O	C	F	B	A	S
N	**224**	0	0	83	38	4	1
O	0	**1**	0	13	5	0	0
C	0	0	**8**	9	13	10	0
F	37	0	8	**385**	242	45	0
B	29	0	13	237	**878**	274	10
A	4	0	9	20	302	**771**	59
S	0	0	0	1	6	97	**84**

ベル数には大きく偏りがあるため，学習セット中のラベル数に反比例する重み c を損失関数に付加した．

$$w_c = \sqrt{\frac{\max_{c' \in \mathscr{C}} \mathrm{count}_{c'}}{\mathrm{count}_c}}, \tag{1}$$

ここで $\mathscr{C}$ はラベルの集合を，count_c はラベル c の学習データ中の出現回数を表す．

学習時には Adam [16] を利用し，初期学習率 1e-5 で 30 エポックのミニバッチ学習を行った．ミニバッチサイズは (4, 8, 16)，追加全結合層のドロップアウト率は (0.1, 0.3, 0.5, 0.75) を比較し，開発セットにおける分類精度が最も高くなるものを選択し，文意解釈性予測モデルはミニバッチサイズ 4，ドロップアウト率 0.75，文意正確性予測モデルはミニバッチサイズ 8，ドロップアウト率 0.5 を使用した．

4.3　結果

表 8, 9 に文意解釈性，文意正確性に対する予測の混同行列をそれぞれ示す．

文意解釈性の予測では，テストセットに対する分類精度は 0.578 であり，学習セット，開発セットに対する分類精度はそれぞれ 0.999，0.647 であった．表 8 から分かるように，誤分類の多くは隣接するクラスに対するもので，離れたクラス (F → {A, S}, D → S, A → F, および S → {F, D}) への誤分類率は 0.43% と低い水準であった．付録の表 12 には文意解釈性予測の適合率・再現率・F 値を示している．

文意正確性の予測では，テストセットに対する分類精度は 0.600 であり，学習セット，開発セットに対する分類精度はそれぞれ 0.998，0.632 であった．出現回数の少なかったラベル (O および C) の予測は重み付き学習を行ったものの芳しくなく，これらの予測精度の向上には誤訳例を学習データにより多く含めることが必要であることが示唆される．表 9 からはラベル S と予測されたものの多く (93.5%) は S または A とラベル付けされた良い翻訳結果であることが分かり，一定の有効性が認められる．しかしながら，F と B の間の混同，特に F のものを B と予測してしまう割合がかなり高いことから，より精緻な識別を可能にする手法やモデルが必要であると言える．付録の表 13 には文意正確性予測の適合率・再現率・F 値を示している．

5　おわりに

本稿では，機械翻訳による深刻な誤訳に着目した分類型翻訳評価に向けての本研究のアプローチを示し，それに基づく文意解釈性と文意正確性の評価基準，評価データセットの構築について述べ，また自動評価実験の結果を示した．

評価データセットは WMT の評価尺度タスクのデータに対して人手で文意解釈性と文意正確性のラベルを付与する形で構築した．評価データセットの分析により，従来用いられてきた DA スコアに基づく人手評価は内容の矛盾等の深刻な誤りを含むような訳文よりも無関係な訳文や理解不能な訳文を低く評価しており，深刻な誤訳の識別には十分でないことが明らかになった．また，本データセットを利用した自動評価実験では，文意解釈性・文意正確性ともおよそ 60%の分類精度で評価ラベルの予測ができることを示した．本データセットは別途 Web サイト[3)]で公開予定である．

今後は WMT の評価尺度タスク以外のデータに対しても同様の人手評価を行った評価データセットの構築を予定している．また，ラベルの偏りへの対応としてのデータ拡張，特に誤訳例の拡張が重要な課題と言える．Popović のような語句単位の評価と本研究の文単位の評価の関係も精緻な評価や自動誤り箇所検出のために有用である．さらに，本研究で提案する分類型翻訳評価を NMT 学習時の目的関数に反映させ，深刻な誤訳の生じにくい NMT の実現を目指したい．

謝辞

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

3)　https://github.com/ksudoh/wmt15-17-humaneval

参考文献

[1] 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.

[2] 嶋中宏希, 梶原智之, 小町守. 事前学習された文の分散表現を用いた機械翻訳の自動評価. 自然言語処理, Vol. 26, No. 3, pp. 613–634, 2019.

[3] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*, 2020.

[4] Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. MoverScore: Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance. 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. 563–578, Hong Kong, China, November 2019. Association for Computational Linguistics.

[5] Thibault Sellam, Dipanjan Das, and Ankur Parikh. BLEURT: Learning Robust Metrics for Text Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7881–7892, Online, July 2020. Association for Computational Linguistics.

[6] Yvette Graham, Timothy Baldwin, Meghan Dowling, Maria Eskevich, Teresa Lynn, and Lamia Tounsi. Is all that Glitters in Machine Translation Quality Estimation really Gold? In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 3124–3134, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee.

[7] LDC. Linguistic Data Annotation Specification: Assessment of Fluency and Adequacy in Translations Revision 1.5, January 25, 2005 . Technical report, Linguistic Data Consortium, 2005.

[8] Isao Goto, Bin Lu, Ka Po Chow, Eiichiro Sumita, and Benjamin Tsou. Overview of the Patent Machine Translation Task at the NTCIR-9 Workshop. In *Proceedings of NTCIR-9 Workshop Meeting*, pp. 559–578, 12 2011.

[9] Arle Richard Lommel, Aljoscha Burchardt, and Hans Uszkoreit. Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics. *Tradumàtica: tecnologies de la traducció*, No. 12, pp. 455–463, 12 2014.

[10] Maja Popović. Relations between comprehensibility and adequacy errors in machine translation output. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pp. 256–264, Online, November 2020. Association for Computational Linguistics.

[11] Maja Popović. Informative Manual Evaluation of Machine Translation Output. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 5059–5069, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[12] John S. White, Theresa A. O'Connell, and Francis E. O'Mara. The ARPA MT Evaluation Methodologies: Evolution, Lessons, and Future Approaches. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas*, Columbia, Maryland, USA, October 5-8 1994.

[13] Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. (Meta-) Evaluation of Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pp. 136–158, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[14] Michael Denkowski and Alon Lavie. Choosing the Right Evaluation for Machine Translation: an Examination of Annotator and Automatic Metric Performance on Human Judgment Tasks . In *Proceedings of the Ninth Biennial Conference of AMTA 2010*, 2010.

[15] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint 1907:11692*, 2019.

[16] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *Proceedings of the Third International Conference on Learning Representations*, 2015.

A　人手評価分析

表 10，11 は本稿で述べた評価データセットにおいて各アノテータが付した文意解釈性および文意正確性ラベルの分布を示している．

表 10　各アノテータのラベル分布（文意解釈性）

	A	B	C	Ave.
F	0.098	0.099	0.111	0.103
D	0.167	0.220	0.181	0.189
B	0.356	0.406	0.222	0.328
A	0.124	0.240	0.219	0.195
S	0.254	0.035	0.266	0.185

表 11　各アノテータのラベル分布（文意正確性）

	A	B	C	Ave.
N	0.098	0.099	0.098	0.098
O	0.004	0.001	0.011	0.005
C	0.009	0.019	0.086	0.038
F	0.205	0.187	0.311	0.234
B	0.374	0.343	0.146	0.288
A	0.233	0.296	0.271	0.267
S	0.076	0.005	0.076	0.069

B　自動評価

表 12，13 は文意解釈性および文意正確性の自動評価実験における，ラベル予測の適合率・再現率・F 値を示している．

表 12　文意解釈性予測の適合率，再現率，F 値

	適合率	再現率	F 値
F	0.769	0.730	0.749
D	0.581	0.438	0.499
B	0.613	0.583	0.598
A	0.439	0.623	0.515
S	0.698	0.569	0.627
Ave.	0.620	0.589	0.598

表 13　文意正確性予測の適合率，再現率，F 値

	適合率	再現率	F 値
N	0.762	0.640	0.696
O	1.000	0.053	0.100
C	0.211	0.200	0.205
F	0.515	0.537	0.526
B	0.592	0.609	0.600
A	0.642	0.662	0.652
S	0.545	0.447	0.491
Ave.	0.609	0.450	0.467