異言語間対話を支援する日英雑談対話誤訳検出

李云蒙 [1] 鈴木潤 [1,3] 森下睦 [2] 阿部香央莉 [1]
徳久良子 [1] ブラサール・アナ [3,1] 乾健太郎 [1,3]
[1] 東北大学 [2]NTT コミュニケーション科学基礎研究所 [3] 理化学研究所
li.yunmeng.r1@dc.tohoku.ac.jp

概要

本論文では，現在の機械翻訳手法の限界に起因する誤訳を検出し，異言語間対話を支援するシステムの開発について述べる．システムのベースラインとして誤訳検出器を学習し，またこの評価のために，複数ターンの雑談をもとに構成された日英雑談対話対訳コーパス「**BPersona-chat**」 [1] に対し，機械翻訳文を利用した低品質な翻訳および各翻訳に対するクラウドソーシングによる品質評価を加え，コーパスを再構築した．結果として，ベースライン検出器は簡単な誤訳を検出することができた．このベースラインを基盤とし，より複雑な異言語間対話における誤訳検出器の構築を目指す．

1 はじめに

国際化の進展に伴い，異言語間対話の必要性が高まっている．しかし，現在の機械翻訳技術は，系統だった文書の翻訳では確かな性能を発揮するが [2, 3, 4]，対話翻訳に適しているとは言い難い [5, 6, 7, 8]．機械翻訳システムで誤訳が生成された場合，ユーザーはその誤訳を識別できず，混乱や誤解を招く可能性がある．本研究では，このような誤訳を検出し，その発生をユーザーに通知することで，潜在的な誤解を低減する異言語間対話支援システムの開発を目指す．このシステムの核心として，本論文では，異言語間対話翻訳の誤訳を検出する誤訳検出タスクの提案およびベースラインとして BERT を用いた誤訳検出器の作成を行った．本誤訳検出タスクの説明図を図 1 に示す．この誤訳検出では，異言語間対話において機械翻訳モデルが誤訳文または文脈との関連性が薄い翻訳文を生成した場合，誤訳検出器がユーザーに誤訳の可能性があるという警告を知らせる．この警告メッセージによって，A 言語側（説明図の日本語側）のユーザーはより機械翻訳が翻訳しやすい形にテキストを修正することが可能である．また，B 言語側（説明図の英語側）のユーザーも，不適切または文脈にそぐわない単語や文章が翻訳エラーである可能性が高いことを認知できる．

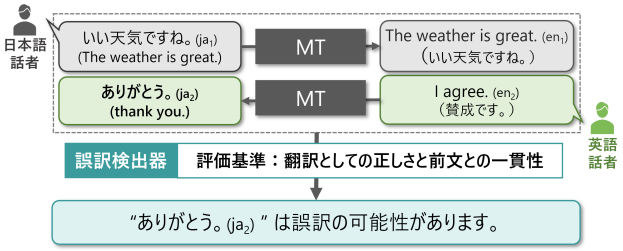

図 1 機械翻訳を介した異言語間対話における誤訳検出タスクの概略図．検出器は翻訳 ja_2 の正確さ・一貫性の評価を行う．

検出した誤訳を評価するために，我々は既存の雑談対訳コーパス「**BPersona-chat**」 [1] に加え，新たに比較的低品質な機械翻訳文およびクラウドソーシングでラベル付けした翻訳の品質ラベル（**正訳**または**誤訳**）を加えたデータセットを作成した．実験では，異なる言語間の雑談対話の翻訳が正しいか誤りであるかを分類する誤訳検出器を学習し（図 1），その性能を作成したデータセットで評価した．本論文の主な貢献は（1）異言語間対話誤訳検出タスクの提案（2）BPersona-chat コーパスの再構築（3）異言語間対話支援システムを開発するためのベースライン誤訳検出器の提供の三つである．

2 誤訳検出タスクの定義

本研究で扱う「異言語間対話における誤訳検出タスク」を定義する．まず，本研究では「異言語間対話」を異なる言語での雑談対話と定義する．したがって，「異言語間対話における誤訳検出タスク」とは，機械翻訳システムを介した異なる言語での対話において，発話（応答）が翻訳システムによって正しく翻訳されたかどうかを予測するタスクとする．

本タスクでは，入力として異言語間対話中の文脈

発話者	原文	人間による翻訳	モデル A の翻訳（低品質）	モデル B の翻訳（高品質）
person 1	hi I am sally, I live with my sweet dogs in taos, new mexico.	こんにちは、サリーです。ニューメキシコ州タオスで愛犬達と一緒に暮らしています。（正訳）	こんにちは私はサリー・ドリー新しい犬とメキシコの新しいメキシコの新しい犬と住んでるの（誤訳）	こんにちは、私はサリー、ニューメキシコ州タオスでかわいい犬たちと暮らしています。（正訳）
person 2	hi! I have just been sitting here playing the piano and singing along	こんにちは！私はここに座ってピアノを弾きながら歌ってたところです。（正訳）	ピアノと歌を歌ってたのよ（誤訳）	私は今ここでピアノを弾きながら歌っています。（誤訳）
person 1	...	...	...	...

表 1　原文に対して三つの翻訳文と人間による評価を付加した BPersona-chat の例.

情報（例えば，直前の相手の発話，相手の発話の翻訳，応答，応答の翻訳など）を受け取り，出力として応答の翻訳が誤りであるかどうかを出力する誤訳検出器を構築し，それを用いて誤訳判定を行うことを想定する．図 1 に，日本語・英語対話における英語応答の日本語訳を評価する例を示す．機械翻訳システムによって，日本語話者の最初の発話 ja_1 が en_1 に翻訳され，英語話者の応答 en_2 が ja_2 に翻訳される．このとき，誤訳検出器は直前の話者の発話，その翻訳と応答 (ja_1, en_1, en_2) を参照し，en_2 の翻訳 ja_2 が正確かつ一貫しているかどうかを予測する．この例では，検出器は「I agree（賛成です）」という応答に対する「ありがとう」という翻訳が誤訳であると評価している．検出器はこのように，会話者に関係なく，各発話の翻訳が行われるたびに，翻訳文を評価して誤訳の有無を検出するものと想定する．

3　関連研究

翻訳品質推定タスク　Wikipedia の記事や Amazon のレビューなど，主に人間の書いた文章を対象とした品質評価タスク [9, 10] と比較して，我々の翻訳品質評価タスクは翻訳文を対象とするという点で新しいタスク設定となる．その上，本タスクでは，対話翻訳の誤訳を検出しするため，対話で用いる口語文の文脈を理解する必要がある．

並列対話コーパス　日英対話コーパスとして，商談シーンを収録した Business Scene Dialog [11] がある．しかし，我々のタスクではより日常的なシーンでの対話支援を目的としている．我々は他言語の並列対話コーパスの構築方法を参考に，タスクに相応しい日英雑談対話コーパスを構築した．ここでは，英独対話コーパス BConTrasT [7] と英中対話コーパス BMELD [8] の構築を参照した．英独対話コーパス BConTrasT は，WMT2020 対話翻訳タスクの主催者が英語対話コーパス Taskmaster-1 [12] をドイツ語に翻訳して作成したコーパスである．また，英中対話コーパス BMELD も同様に，MELD [13] という英語対話コーパスを中国語に翻訳して作成したコーパスである．

これらの並列対話コーパスを参考に，我々は雑談対話が含まれる英語対話コーパス Persona-chat [14] と日本語対話コーパス JPersona-chat [15] を元になるデータセットとして採用し，日英雑談対話対訳コーパス BPersona-chat を構築した [1]．品質を保証するため，クラウドワーカーに依頼して Persona-chat 中にある不自然な話題転換や誤解などが含まれていて一貫性がない対話をフィルタリングし，プロの翻訳家の手で翻訳を付与した．

本研究では、提案タスクを評価するため，一度構築した BPersona-chat に新たな翻訳とアノテーションを付与し、評価用データセットとして再構築した．

4　評価用データセット

本タスクの評価用データセットにするため，BPersona-chat に新たな翻訳文を追加した．さらに，翻訳が対話翻訳として許容できるかについて評価し，データセットを再構築した．詳細は以下の通りである．

4.1　対話データの人間による翻訳

BPersona-chat を構築する際，Persona-chat の合計 2,940 の発話から上位 200 件の発話と，JPersona-chat の合計 2,740 の発話から上位 250 件の発話を抽出し，それぞれの対象言語に翻訳した[1)]．翻訳者には，翻訳の正確さと対話としての一貫性の双方に配慮してもらった．その結果，450 の対話（5,680 の発話）とその翻訳を得ることができた [1]．BPersona-chat を再構築する際，上記のデータも人間による翻訳として活用した．

1)　JPersona-chat を翻訳することは著者から同意を得た．

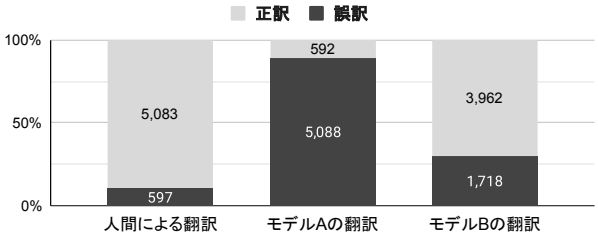

図 2 人間による翻訳，モデル A の翻訳，モデル B の翻訳それぞれにおけるアノテーション後の正訳と誤訳の割合．

4.2 対話データの機械翻訳

誤訳検出タスクを解く検出器のベースラインを構築するため，BPersona-chat における専門家の翻訳に加え，主に負例として利用することを念頭に機械翻訳で生成した翻訳結果を用意した．機械翻訳結果を得るために，２種類の機械翻訳モデルを用意した．その１つは，OpenSubtitles2018 [16] によって学習した Transformer ベースのニューラル機械翻訳（NMT）であり，これをモデル A と呼ぶ．モデル A の出力を BPersona-chat で評価すると，BLEU の値 [17] は 4.9 となった[2)]．なお，モデル A の BLEU 値が比較的低い原因は，OpenSubtitles2018 と BPersona-chat のドメインが違うことによるものと考えられる．しかし，データセットを構築するためにあえて低品質の翻訳結果が必要であったため，これは望ましい結果である．次に，より良い機械翻訳文を生成するため，公開されている機械翻訳 API（DeepL API）を使用した．これをモデル B と呼ぶ．BPersona-chat をモデル B で評価すると BLEU スコアは 26.4 となった．

4.3 翻訳の人間による評価

人間の翻訳および機械翻訳モデル A,B が生成した翻訳候補に対し，誤訳検出タスクにおける翻訳品質ラベルを付与した．なお，この判定は，英語と日本語の両方に堪能なクラウドワーカーを集めて実施した．図 2 に示す結果から，モデル A の翻訳の 89.58%，モデル B の翻訳の 30.25%，人間の翻訳の 10.51% が誤訳と評価された[3)]．

発話に付与された翻訳が全て正しくない場合，誤訳検出タスクにおいて参照する文脈として使用することは難しい．したがって，人間による翻訳，モデル A の翻訳，モデル B の翻訳の全てが誤っている 159 件の発話は評価データとして用いないことと

2） NMT モデル A の学習は付録 A を参照．
3） クラウドソーシングの詳細については付録 C を参照．

	日→英	英→日
マジョリティクラス分類器	56.94	57.54
マイノリティクラス分類器	43.06	42.46
誤訳検出器	76.27	77.06

表 2 マジョリティクラス分類器，マイノリティクラス分類器，および誤訳検出器の精度．

	日→英			英→日		
	F	**(Pre**	**Rec)**	**F**	**(Pre**	**Rec)**
誤訳検出器	73.30	(71.10	75.65)	75.03	(69.75	81.18)

表 3 BPersona-chat における誤訳検出器の F 値，適合率と再現率．

した．最終的には，英語発話 2,674 件に対して日本語訳 8,022 件が得られ，そのうち 3,406 件が誤訳，残りの 4,616 件が正訳であった．また，日本語発話 2,397 件に対して英語訳 7,190 件が得られ，3,096 件が誤訳，4,094 件が正訳となった．これらをまとめて BPersona-chat コーパスを再構築した[4)]．表 1 は BPersona-chat のサンプルである．

5 ベースライン誤訳検出器

誤訳検出器のベースラインとして，BERT に基づく二値分類モデル [18, 19] を学習・評価した[5)]．英語発話 en_2 の日本語翻訳 ja_2 の誤訳判定をする場合，分類モデルの入力は “ja_1[SEP]en_1[SEP]en_2[SEP]ja_2” とした．同様に，日本語発話 ja_2 の翻訳 en_2 の誤訳判定をする場合，入力を “en_1[SEP]ja_1[SEP]sja_2[SEP]en_2” とした[6)]．BERT の元実験と同様，分類結果に SoftMax 関数を適用して予測値を計算した．

分類モデルの学習には OpenSubtitles2018 データセット約 100 万の発話を用いた．参照訳を正例，低品質翻訳モデル A による出力を（4.2 節）で擬似的な負例とし，HuggingFace[7)]が提供する多言語 BERT モデルを再学習して，英語から日本語，日本語から英語両方向の誤訳検出器を構築した．

6 実験

本節では，異言語間対話訳における誤訳検出器の試行と評価結果を報告する．

4） https://github.com/cl-tohoku/BPersona-chat
5） 分類モデルの学習については付録 B を参照．
6） 各文脈情報の区別を示すために [SEP] を，データの先頭を示すために [CLS] を，パディングトークンとして [PAD] を使用した．
7） https://huggingface.co/

日→英								
人間による翻訳			モデル A の翻訳			モデル B の翻訳		
	正訳	誤訳		正訳	誤訳		正訳	誤訳
正訳	1879	207	正訳	11	155	正訳	1252	590
誤訳	290	21	誤訳	90	2140	誤訳	374	181
英→日								
人間による翻訳			モデル A の翻訳			モデル B の翻訳		
	正訳	誤訳		正訳	誤訳		正訳	誤訳
正訳	2406	176	正訳	6	265	正訳	1005	758
誤訳	83	9	誤訳	53	2350	誤訳	505	406

表 4　BPersona-chat に対する誤訳検出器の混同行列（行ヘッダが人間による評価，列ヘッダが検出器による予測）．

6.1　評価指標

マジョリティクラスとマイノリティクラス分類器　誤訳検出器の予測が偶然正解したものではないことを確認するために，マジョリティクラス分類器，マイノリティクラス分類器，そして誤訳検出器の精度を計算した．

F 値，適合率と再現率　誤訳検出器の性能は F 値（**F**）によって評価した．また，適合率（**Pre**）と再現率（**Rec**）も合わせて計算した．このとき真値（T）を誤訳ラベルが付いた事例，正例（P）を検出器が誤訳と判定した事例とした．

混同行列　誤訳検出器の性能を評価するため，翻訳が人間，NMT モデル A，NMT モデル B のいずれによって翻訳されたかに応じて混同行列を示す．

6.2　結果

誤訳検出器は異言語間対話中の誤訳をある程度分類することができることがわかった．表 2 に示す精度から，誤訳検出器はマジョリティクラスとマイノリティクラス分類器と比較して高い性能を得た．この結果は，検出器は片方の結果のみを偏って出力しているわけではないということを示している．さらに，表 3 に示す F 値，適合率，再現率によると，誤訳検出器は BPersona-chat の誤訳をそれなりの精度で識別することが可能であった．

しかし，表 4 に翻訳の種類に応じた混同行列より，検出器は高品質なモデル B で生成された翻訳をうまく識別することはできなかった。．検出器はモデル B で生成された誤訳の半分以上を正訳として判断した．うまく判別できなかった理由として，誤訳検出器が，低品質なモデル A で生成された負例で学習を行なっていることが挙げられる．

誤訳検出器と従来の BLEU 計算による評価を比較するため，BPersona-chat 中の誤訳文に対する sentence-BLEU スコア [20][8] を計算した．sentence-BLEU を計算する際，正訳ラベルがついた翻訳文を参照訳とした．その結果，誤訳ラベルがついた例のうち，誤訳検出器によって誤訳と判断され，かつ sentence-BLEU スコアも 60 以下となった翻訳文は日本語誤訳の 78.83%，英語誤訳の 76.75% となった．なお，誤訳ラベルがついた例のうち，誤訳検出器では正しく誤訳と判断されたが sentence-BLEU スコアが 60 以上となる翻訳文は，日本語誤訳の 2.14%，英語誤訳の 21.83% となった。結果として，誤訳検出器は正しい翻訳（参照訳）の有無に依存せず，sentence-BLEU スコアを基準として誤訳を判別したときの精度の 70〜80 %程度に匹敵すると判明した．また，特に英語訳文を判定する際，ベースライン誤訳検出器は sentence-BLEU スコアが高い，すなわち表層的には正訳と類似している誤訳文も検出可能であることを示唆する結果が得られた．このように，ここに提案する誤訳検出器は，BLEU のような指標では判別できない誤訳を判別できる可能性が高く，この結果は誤訳検出器が異言語間対話の支援システムとして活用できる可能性を示している．

7　まとめ

本論文では，異言語間でのコミュニケーションを支援するために，異言語間対話における誤訳検出タスクを提案した．本研究における評価のために，複数ターンの雑談をもとに構成された日英対訳コーパスに，比較的低品質な機械翻訳文およびクラウドソーシングによる翻訳の品質（正訳または誤訳）の分類を付与した評価用データセットを構築した．また，誤訳を検出する誤訳検出器を学習し，異言語対話における誤訳の検出を支援するシステムのベースラインを構築した．更に，上記の評価用データセットを用いてベースラインを評価した．

今後，異言語間対話支援システムの向上に向けて，より詳細に誤訳の可能性を示す機能の実現を目指している．発展として，翻訳に含まれる具体的な誤りを特定できるように，二値分類を複数ラベルに改良することを考えている．また，発話候補を参考情報として提供し，ユーザーに発話の修正を促すことも検討したい．

8）https://www.nltk.org/_modules/nltk/translate/bleu_score.html

謝辞

本研究は JST 科学技術イノベーション創出に向けた大学フェローシップ創設事業 JPMJFS2102，JST CREST Grant Number JPMJCR20D2，JST ムーンショット型研究開発事業 JPMJMS2011，JSPS 科研費 JP20J21694 の支援を受けたものである．

BPersona-chat の配布を許可いただいた Persona-chat [14] の開発者様と JPersona-chat [15] の開発者様に深く感謝申し上げます．

Amazon Mechanical Turk（https://www.mturk.com/）と Crowdworks（https://crowdworks.jp/）においてクラウドワーカーとしてご協力いただいた皆様へ，深く感謝を申し上げます．

本研究を進めるにあたり，頻繁に議論に参加していただいた東北大学 Tohoku NLP グループの皆様へ感謝いたします．

参考文献

[1] Yunmeng Li, Jun Suzuki, Makoto Morishita, Kaori Abe, Ryoko Tokuhisa, Brassard Ana, and Inui Kentaro. Bpersona-chat: A coherence-filtered english-japanese dialogue corpus. In **Proceedings of NLP2022**, pp. E7–3, 2022.

[2] Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. Findings of the 2019 conference on machine translation (WMT19). In **Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)**, pp. 1–61, 2019.

[3] Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. Findings of the 2020 conference on machine translation (WMT20). In **Proceedings of the Fifth Conference on Machine Translation**, pp. 1–55, 2020.

[4] Toshiaki Nakazawa, Nobushige Doi, Shohei Higashiyama, Chenchen Ding, Raj Dabre, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Yusuke Oda, Shantipriya Parida, Ondřej Bojar, and Sadao Kurohashi. Overview of the 6th workshop on Asian translation. In **Proceedings of the 6th Workshop on Asian Translation**, pp. 1–35, 2019.

[5] Samuel Läubli, Rico Sennrich, and Martin Volk. Has machine translation achieved human parity? a case for document-level evaluation. In **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**, pp. 4791–4796, 2018.

[6] Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. Attaining the unattainable? reassessing claims of human parity in neural machine translation. In **Proceedings of the Third Conference on Machine Translation: Research Papers**, pp. 113–123, 2018.

[7] M. Amin Farajian, António V. Lopes, André F. T. Martins, Sameen Maruf, and Gholamreza Haffari. Findings of the WMT 2020 shared task on chat translation. In **Proceedings of the Fifth Conference on Machine Translation**, pp. 65–75, 2020.

[8] Yunlong Liang, Fandong Meng, Yufeng Chen, Jinan Xu, and Jie Zhou. Modeling bilingual conversational characteristics for neural chat translation, 2021.

[9] Lucia Specia, Frédéric Blain, Marina Fomicheva, Erick Fonseca, Vishrav Chaudhary, Francisco Guzmán, and André F. T. Martins. Findings of the WMT 2020 shared task on quality estimation. In **Proceedings of the Fifth Conference on Machine Translation**, pp. 743–764, 2020.

[10] Erick Fonseca, Lisa Yankovskaya, André F. T. Martins, Mark Fishel, and Christian Federmann. Findings of the WMT 2019 shared tasks on quality estimation. In **Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)**, pp. 1–10, 2019.

[11] Matīss Rikters, Ryokan Ri, Tong Li, and Toshiaki Nakazawa. Designing the business conversation corpus. In **Proceedings of the 6th Workshop on Asian Translation**, pp. 54–61, Hong Kong, China, November 2019. Association for Computational Linguistics.

[12] Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Daniel Duckworth, Semih Yavuz, Ben Goodrich, Amit Dubey, Andy Cedilnik, and Kyu-Young Kim. Taskmaster-1: Toward a realistic and diverse dialog dataset, 2019.

[13] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations, 2019.

[14] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too?, 2018.

[15] Hiroaki Sugiyama, Masahiro Mizukami, Tsunehiro Arimoto, Hiromi Narimatsu, Yuya Chiba, Hideharu Nakajima, and Toyomi Meguro. Empirical analysis of training strategies of transformer-based japanese chit-chat systems, 2021.

[16] Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In **Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)**, pp. 1742–1748, 2018.

[17] 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.

[18] 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, 2019.

[19] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations**, pp. 38–45, 2020.

[20] Steven Bird, Ewan Klein, and Edward Loper. **Natural language processing with Python: analyzing text with the natural language toolkit**. ” O’Reilly Media, Inc.”, 2009.

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In **Advances in Neural Information Processing Systems 31 (NIPS 2017)**, pp. 5998–6008, 2017.

[22] Jörg Tiedemann and Yves Scherrer. Neural machine translation with extended context. In **Proceedings of the Third Workshop on Discourse in Machine Translation**, pp. 82–92, 2017.

[23] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In **Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)**, 2015.

[24] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. **Journal of Machine Learning Research**, Vol. 15, No. 56, pp. 1929–1958, 2014.

[25] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. In **2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016)**, pp. 2818–2826, 2016.

[26] Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. Scaling Neural Machine Translation. In **Proceedings of the Third Conference on Machine Translation (WMT 2018)**, pp. 1–9, 2018.

[27] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google’s neural machine translation system: Bridging the gap between human and machine translation. **CoRR**, Vol. abs/1609.08144, , 2016.

[28] 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.

[29] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In **Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics.

Architecture	2-to-2 Transformer [21, 22]
Enc-Dec layers	6
Attention heads	8
Word-embedding dimension	512
Feed-forward dimension	2,048
Share all embeddings	True
Optimizer	Adam ($\beta_1 = 0.9, \beta_2 = 0.98, \epsilon = 1 \times 10^{-8}$) [23]
Learning rate schedule	Inverse square root decay
Warmup steps	4,000
Max learning rate	0.001
Initial Learning Rate	1e-07
Dropout	0.3 [24]
Label smoothing	$\epsilon_{ls} = 0.1$ [25]
Mini-batch size	8,000 tokens [26]
Number of epochs	20
Averaging	Save checkpoint for every 5000 iterations and take an average of last five checkpoints
Beam size	6 with length normalization [27]
Implementation	`fairseq` [28]

表 5 NMT モデル A 学習のハイパーパラメータ一覧

Architecture	BERT (base) [18]
Optimizer	Adam ($\beta_1 = 0.9, \beta_2 = 0.98, \epsilon = 1 \times 10^{-8}$, weight decay=0.01) [23]
Learning rate schedule	Inverse square root decay
Max learning rate	0.001
Mini-batch size	16 samples
Number of epochs	1
Implementation	`transformers` [19]

表 6 分類モデル学習のハイパーパラメータ一覧

A 翻訳モデルの学習設定

ニューラル機械翻訳モデル A を学習する際，まずは BPE [29] でコーパスをトークナイズしてサブワードにする．語彙の大きさは 32,000 とした．文脈を考慮するために，2 つの入力文を与えて 2 つ連続して出力する 2-to-2 Transformer-based NMT モデル A [22] を学習した．表 5 にハイパーパラメータの一覧を示している．

B 分類モデルの学習設定

分類モデルの学習について説明する．表 6 にハイパーパラメータの一覧を示している．

C クラウドソーシング関連設定

C.1 Persona-chat のフィルタリング

Amazon Mechanical Turk（https://requester.mturk.com/）のクラウドワーカーに，Persona-chat の一貫性がない対話をフィルタリングするよう依頼した．以下に当てはまる場合，「一貫性がない」対話であると定義した．

- questions being ignored;
- the presence of unnatural topic changes;
- one is not addressing what the other said;
- responses seeming out of order;
- or being hard to follow in general.

ワーカーには，誤字脱字などの細かいことは気にせず，大まかな流れを把握するよう指示した．

フルラウンドでは，Persona-chat から 1,500 の対話を選択した．クラウドワーカーは，予選を経て選ばれた．各クラウドワーカーは対話 5 件を評価し，各対話は異なるワーカー 10 人によって評価された．選ばれた 200 対話はワーカー 10 人のうち，少なくとも七人が正しいかつ一貫していると評価した対話となった．

C.2 対話に対する翻訳の評価

クラウドワークス（https://crowdworks.jp/）のクラウドワーカーに，BPersona-chat の人間による翻訳と機械翻訳を低品質か高品質かラベル付けしてもらうタスクを行った．クラウドワーカーの資格は，日本語はネイティブレベル，英語はビジネス・アカデミックレベルに到達できるレベルとなっている．このタスクでは，以下の場合に「低品質である」と定義した．

- the translation is incorrect;
- parts of the source chat are lost;
- there are serious grammatical or spelling errors that interfere with understanding;
- the person's speaking style changes from the past utterance;
- the translation is meaningless or incomprehensible;
- or the translation is terrible in general.

ワーカーは，対話全体が含まれるファイルを一つ一つ確認し，各発話を評価することができた．

ワーカーによって，クラウドワーカーは二週間でファイル 50 個から 300 個程度を評価した．このタスクは事前に予選を行い，予選を通過したワーカーのみ本選（実際の作業）に参加可能とした．