異言語間対話支援における誤訳警告メッセージの有効性調査

Yunmeng Li[1] 鈴木潤[1,3] 森下睦[2] 阿部香央莉[3,1*] 乾健太郎[4,1,3]
[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 [4]MBZUAI
li.yunmeng.r1@dc.tohoku.ac.jp

概要

現状の機械翻訳システムは，実用レベルに達したと言われるようになったが，異言語間対話支援に利用する場面では，ユーザの意図を正しく翻訳できるかという観点で，まだ実用レベルとは言い難い．誤訳を多く含む異言語間対話支援において，機械翻訳システムを利用する実用的なアプローチの一つは，誤訳に関する警告メッセージを提示し，ユーザの混乱を軽減する方法である．しかし，このような警告メッセージがユーザにどのように受け入れられ，また，どのような利益をもたらすのかについては未検証課題である．本研究では，この課題に取り組み，異言語間対話支援における誤訳警告メッセージの有効性を検証する．

1 はじめに

国際的な交流の需要が年々高まっている昨今，Google Translate[1)]や DeepL[2)]などの機械翻訳アプリケーションが活躍する場面も多くなっている．また，WeChat や LINE などの対話アプリケーションにも，異言語間のコミュニケーションを促進するための翻訳機能が組み込まれている．更に，UD Talk[3)]や Hi Translate[4)]などのプラグイン翻訳アプリケーションも，オンラインコミュニケーションの進化と共に広まっている．

機械翻訳技術は，ニューラル機械翻訳 [1, 2, 3] の急速な発展に伴って，書き言葉の翻訳において頑健な性能を示している [4, 5, 6]．一方で，現行の手法でも，話し言葉や日常会話の翻訳において話者の意図を正しく翻訳できているかという観点で必ずしも十分ではないことが指摘されている [7]．特に異言語間の対話によるコミュニケーションでは，システムが誤訳を生成すると，他の言語を理解できない話し相手が誤りに気づかず，誤解を引き起こして，コミュニケーションを妨害する可能性がある [8, 9]．

対話特有の性質により，対話翻訳において「正解」の定義が複雑であり [10, 11, 9, 12]，誤訳のない対話翻訳システムの実現を目指すことは非現実的である．その代わりに，混乱を減らすために誤訳の可能性を警告メッセージとして提示することで，翻訳システムを強化し，異言語間対話を支援するアプローチが現実的と考えられる．しかし，このような警告メッセージの認識と効果は，これまで検証されておらず，未だに不明確である．

本研究では，この不明確な点を明らかにするために，異言語間対話を支援する機械翻訳の誤訳警告メッセージを提供した場合，警告メッセージが異言語間コミュニケーションにどのように役立つかを人手による主観評価 (アンケート調査) により検証する．アンケートデザインを図 1 に示す．参加者は 3 つの選択肢の中から最も妥当な回答を選び，模擬的な異言語間対話シナリオに参加する．誤訳が発生するたびに，警告メッセージが表示される．対話の最後には，参加者は警告メッセージに対する認識に関する主観評価の質問に答える．具体的には，クラウドソーシングを通じて人手による主観評価を収集した．その結果，(1) 警告メッセージは異言語間対話支援に有用であり，(2) ユーザの対話行動に関して行動変容を促す可能性があることが示された．

この調査は，警告メッセージの実用性を裏付けるものであり，ユーザがどのようなことを警告メッセージに期待しているか，ユーザに効果的な警告メッセージのあり方を示唆する結果となっている．本研究は，異言語間対話支援における機械翻訳の誤訳警告メッセージがユーザに与える影響を初めて調査したものであり，この結果は，より円滑な異言語間対話支援の研究に寄与する貴重な知見を提供する．

---

* 現在はマシンラーニング・ソリューションズ株式会社に所属．
1) https://translate.google.com/
2) https://www.deepl.com/translator
3) https://udtalk.jp/
4) https://bit.ly/3pWhz9T

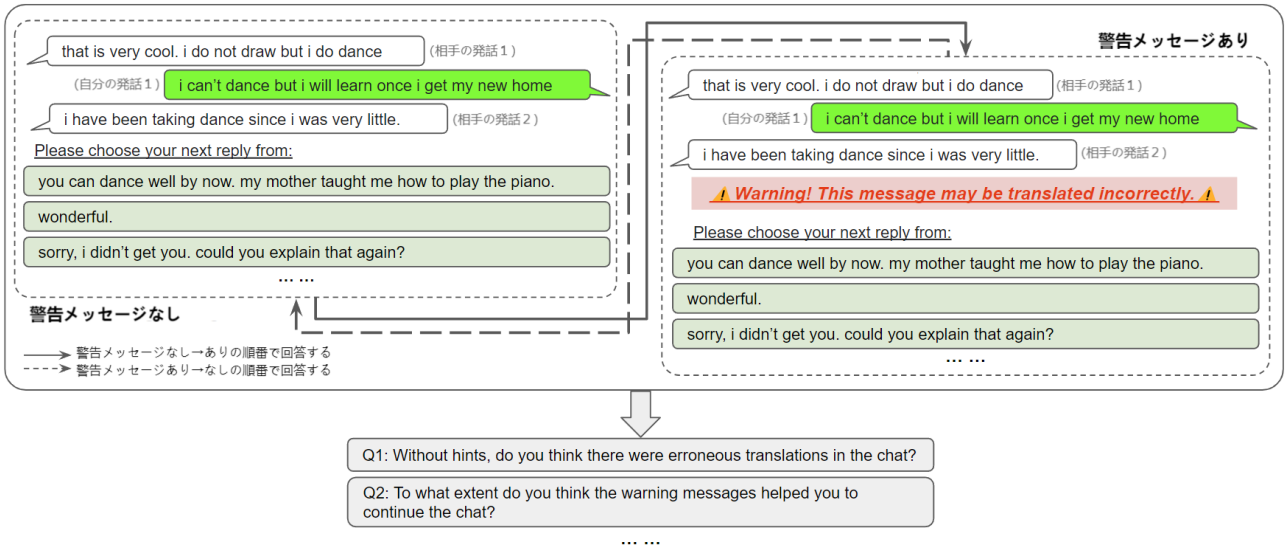

図 1　アンケート調査の説明図．参加者は，警告メッセージなし（左）と警告メッセージあり（右）の 2 ラウンドの対話に参加する．どちらのラウンドでも，内容と回答の選択肢は同じである．2 つのラウンドの順番，すなわち「警告メッセージあり→なし」（実線）か「警告メッセージなし→あり」（点線）かは，参加者にランダムに割り当てられる．

2　関連研究

先行研究においては，不完全ながらも対話に機械翻訳を導入することで異言語間コミュニケーションをより活性化できるという利点が報告されている [7]．一部の研究者は，対話翻訳の表現を向上させるためにモデルを訓練してきた [11, 8, 9]．しかし，曖昧性，省略，複数話者など対話ドメイン独自の特徴によって，対話で翻訳精度を向上させることは困難な課題となっている [10, 9, 12]．これら既存の研究に対し我々は，機械翻訳の精度を 100 %にするのは困難であるとの前提に立ち，対話翻訳のユーザエクスペリエンスを向上させるためのアプローチに焦点を当て，誤訳の可能性をユーザに示唆する警告メッセージを提示することを提案する．翻訳誤りについては近年対話翻訳誤訳検出器の研究もあり，対話翻訳の一貫性と正確性の観点から評価をされている [13]．誤訳検出器の予測が警告メッセージに変換される場合，我々の調査は誤訳検出器の実用的な有効性を評価するのに役立つ可能性がある．

3　調査デザイン

異言語間対話のシミュレーション　誤訳によっての混乱を軽減するため，我々は警告メッセージの提示を提案する．本研究では，その警告メッセージの有効性を調査するため，クラウドソーシングを通じて人手による主観評価を収集した．図 1 に，評価を収集するためのアンケート調査のデザインを示す．

動的なリアルタイム対話は制御が難しく，実験のコストも高い．そこで，本研究では Persona-chat [14] の対話データを基に，ユーザが対話アプリ上で機械翻訳を介して外国人パートナーと対話する状況をシミュレートすることにした．調査の参加者には，はじめに対話の先行文脈として直前の 3 発話（相手の発話 1，自分の発話 1，相手の発話 2）が提示される．参加者は，次の自分の発話として，提示された 3 つの選択肢から最も文脈に合った応答を選択する．この時，直前の相手の発話または送信した自分の発話に対して，誤訳の可能性が高い場合に警告メッセージを提示することを考える．実験の目的は，この警告メッセージの有無によって参加者の反応がどのように変化するかを調査することである．

本研究では，対話参加者は相手言語に習熟していないユーザであると仮定する．この想定をシミュレートするため，参加者には，参加者自身の応答，相手発話の機械翻訳結果，およびそれに対する警告メッセージだけを提示して調査を行った．すなわち，調査内のすべてのテキストは，参加者の母語で提示されている．

対話データのフィルタリング　対話データの質を保証するために，Amazon Mechanical Turk のクラウドソーシングを通じて，Persona-chat から一貫性がない対話を排除した．ここでの「一貫性がない」とは，質問が無視される，不自然な話題の切り替えがある，相手の発言に対処していない，応答が順序から外れているように見える，または一般的に対話の

流れをフォローするのが困難であるものとした．クラウドワーカーの回答に基づいて各対話を採点し，1,500 個の対話のうち，10 人のうち少なくとも 7 人が一貫しているとマークした 200 件を獲得した．この 200 件のうち，6 件を異言語間対話のシミュレーションのベースとして使用した．

誤訳データ　誤りを含む翻訳文（誤訳文）は次の手順で収集した．まず，英日対話翻訳評価データセット BPersona-chat [15] を BLEU スコア [16] が 4.9 の低品質機械翻訳モデル[5]を使って相手発話を翻訳した．次に，得られた訳文に対し，Google 翻訳による別言語への翻訳とターゲット言語への逆翻訳を繰り返し（実験では 20 回），元文からさらに逸脱した訳文を収集した．最後に，得られた訳文から文法誤りや不適格な内容等の誤りが含まれているものを手作業で選んだ．

警告メッセージ　警告メッセージは，対話での誤訳をアンケート調査の参加者に通知するように設計した．図 1 に示したように，対話文が誤訳である場合，参加者に対して誤訳を提示する警告メッセージが表示される．対話では受信と送信の両方が不可欠であるため，警告メッセージを 2 つのタイプに分類した．一方は，受信したメッセージの誤訳を警告するタイプであり，もう一方は、最後に送信したメッセージの誤訳の可能性を示すタイプである[6]．

アンケート質問　対話シミュレーションの後に，警告メッセージに対する参加者の評価を収集するためのアンケートを行う．参加者はまず警告メッセージなしで誤訳に気づいたかどうかを答えるよう求められる．参加者が「はい」と回答した場合，2 つのリッカート尺度 [17, 18] の設問に進む．最初の設問では，「誤訳がどのくらい対話を続けることを妨げられたか」を評価する．2 番目の設問では「誤訳が具体的にどの位置にあったかを把握できたか」尋ねる．参加者は 2 つの設問に対し 1〜5（数字が大きいほど誤訳に対する認識や理解が高い）で評価する．

さらに，参加者は警告メッセージが対話の続行に役立ったと思う程度を評価し，警告メッセージに追加して欲しい機能をにチェックを入れる．選択可能な機能は以下の通り：翻訳の正答率を示す，翻訳候補を表示する，翻訳の誤りを具体的に示す，相手の気持ちを表示する機能．[7]

5）　低品質機械翻訳モデルの学習は付録 A を参照．
6）　図 1 は前者の例である．
7）　特に欲しい機能がなければ空欄での回答も可能とした．

4　クラウドソーシング実験

調査用のアンケートは英語，中国語，日本語で作成し，言語間の違いを観察した．アンケートでの対話データは，英語が堪能な中国語（又は日本語）の母語話者により母語に翻訳し，品質を確保した．警告メッセージの 2 種類ごとに対話を 3 セットを用意し，全部で 6 セットの対話を提供した．参加者は，(1) 対話相手は母国語以外の言語で話しかけている前提であること，(2) 機械翻訳システムが相手のメッセージを翻訳し，対話はユーザの母国語でのみ表示されること，(3) 対話ログを読み，3 つの選択肢のうち最も妥当なものを選ぶこと，(4) 送られてきたメッセージは奇数行目に，自分の答えは偶数行目に表示されることを認識するよう指導した．

回答順番による影響を最小限にするため，参加者に対し，警告メッセージなし→ありの順番で回答するか，警告メッセージあり→なしの順番で回答するかをランダムに割り振りした．このとき，どちらの順番になるかは参加者には知らせずに調査を行った．警告メッセージが表示されるラウンドでは，参加者に警告メッセージの役割を説明し，応答選択の際に参考にできることを伝えた．

各セットに対して少なくとも 50 人の参加者を募り，1 人には同じ対話が 2 回以上回答しないようにした．各言語の調査は Amazon Mechanical Turk （英語），WenJuanXing（中国語），CrowdWorks（日本語）でそれぞれ実施した．調査の参加者には結果を学術目的で使用することを通知した．

5　結果と分析

アンケート調査の結果，最終的に英語では 604 件，中国語では 635 件，日本語では 621 件の回答が集まった．3 言語において，約 70 %の参加者が警告メッセージを「4-役に立った」以上と評価した[8]．

警告メッセージの有無　表 1 により，誤訳に気づいた割合は，アンケートのデザインにおける回答の順番（警告メッセージを含むラウンドを先に行うか否か）に関わらず一貫している．さらに，誤訳に気づいたほとんどの参加者は，その誤訳によって対話の進みが妨害されたと見なした．

注目すべき点として，英語と中国語の結果は比較的類似しているが，日本語の結果の傾向は若干異なっている．特に警告メッセージのない誤訳を認識

8）　全体の結果は付録 B を参照．

表 1　警告メッセージがない場合に誤訳に気づくかどうかの質問についての結果．誤訳に気づいた参加者は，誤訳をどの程度対話を妨害したと考えたかを評価し続けた．

	警告メッセージなしを先に回答		警告メッセージありを先に回答	
	警告メッセーなしで誤訳に気づいた	誤訳は対話の進みを妨害したと感じた	警告メッセーなしで誤訳に気づいた	誤訳は対話の進みを妨害したと感じた
英語	77.2% (234)	70.5% (165)	77.4% (234)	67.1% (157)
中国語	70.2% (228)	72.4% (165)	77.7% (241)	62.7% (151)
日本語	54.5% (175)	69.7% (122)	52.7% (158)	70.3% (111)

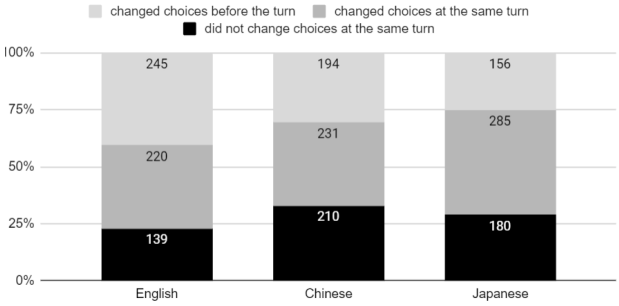

図 2　警告メッセージによって参加者の行動が変化した割合．

できたかどうかという点について，英語や中国語による調査よりも日本語による調査での結果が顕著に低い結果となった．この結果は，日本語の言語的特異性，特に省略の多さに関連している可能性がある．日本語によるアンケート調査の参加者は，主語や目的語が省略されることが多い日本語に合わせて，誤訳を省略の一種であると考えてしまったと見られる．警告メッセージは，その表現が省略されたのではなく誤りであることを日本語話者に理解してもらうのに役立ったと考えられる．

ユーザの行動に与える影響　警告メッセージに関連した選択を分析し，3 つのケースに分類した：(1) 警告メッセージの有無に関わらず選択を変えなかった (2) 警告メッセージにより選択を変えた (3) 警告メッセージが表示される以前に選択を変えた．(1) は参加者が警告メッセージの影響を受けなかったことを示し，(2) は参加者が影響を受けたことを示している．また，(3) は間接的に警告メッセージの影響を受けていると考えられる．なぜなら，警告メッセージを表示するラウンドでは「警告がある場合，誤訳がある（ない場合は誤訳が含まれない）」という指示により，参加者が「誤訳がない」という情報も用いることができるためである．図 2 より，約 75 %の参加者が警告メッセージによって直接的または間接的に選択を変えたということが分かる．

受信時または送信時における警告メッセージの有用性　対話文の誤訳を受信時または送信時に指摘するかどうかに関わらず，3 言語全てで 60 %以上の参加者が警告メッセージが役に立ったと回答した[9]．

警告メッセージに期待される機能　中国語および日本語による参加者は，誤訳が具体的に発生している所を示す機能に対して期待を示した．中国語圏の参加者は誤訳応答を解釈し直す必要があるかどうかを判断するため，正解率も知りたがっている．一方，日本語圏の参加者は他の翻訳案を参考にすることを検討している．英語圏の参加者はすべての機能に対し平均的に投票したが，相手の感情を示す機能に投票した参加者の人数は他言語と比較すると少なかった．まとめると，警告メッセージを強化するためには，誤りが発生している所を示すことに焦点を当てた方が良い可能性があるということが判った[10]．

6　おわりに

我々は，異言語間対話を支援するため，対話における誤訳を提示する警告メッセージの有効性を評価する調査を実施した．クラウドソーシングを通じて，対話誤訳警告メッセージに対する人手による主観評価を収集し，その回答により警告メッセージが有益であるとの結論に至った．警告メッセージの有無に基づいて参加者の選択を比較することにより，警告メッセージが参加者の行動に影響を与えることが明らかになった．また，参加者は警告メッセージに対して，(1) 翻訳の具体的な誤りを示すこと，(2) 正解率を提示すこと，(3) 代替翻訳の提案を行うことを期待していることが示された．

9)　具体的な割合は付録 B を参照．
10)　具体的な結果は付録 B を参照．

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 JPMJMS2011-35 の支援を受けたものである．Amazon Mechanical Turk（https://www.mturk.com/）, WenJuanXing（https://www.wjx.cn/）と Crowdworks（https://crowdworks.jp/）においてご協力いただいた皆様へ感謝を申し上げます．研究を進めるにあたり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] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. **arXiv preprint arXiv:1409.0473**, 2014.

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. **Advances in neural information processing systems**, Vol. 30, , 2017.

[3] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. Convolutional sequence to sequence learning. In **International conference on machine learning**, pp. 1243–1252. PMLR, 2017.

[4] Shuming Ma, Dongdong Zhang, and Ming Zhou. A simple and effective unified encoder for document-level machine translation. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 3505–3511, Online, July 2020. Association for Computational Linguistics.

[5] Sameen Maruf and Gholamreza Haffari. Document context neural machine translation with memory networks. In **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 1275–1284, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[6] 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, Hong Kong, China, November 2019. Association for Computational Linguistics.

[7] David C Uthus and David W Aha. Multiparticipant chat analysis: A survey. **Artificial Intelligence**, Vol. 199, pp. 106–121, 2013.

[8] 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, Online, November 2020. Association for Computational Linguistics.

[9] Yunlong Liang, Fandong Meng, Yufeng Chen, Jinan Xu, and Jie Zhou. Modeling bilingual conversational characteristics for neural chat translation. In **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pp. 5711–5724, Online, August 2021. Association for Computational Linguistics.

[10] 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, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[11] Sameen Maruf, André F. T. Martins, and Gholamreza Haffari. Contextual neural model for translating bilingual multi-speaker conversations. In **Proceedings of the Third Conference on Machine Translation: Research Papers**, pp. 101–112, Brussels, Belgium, October 2018. Association for Computational Linguistics.

[12] Yunlong Liang, Chulun Zhou, Fandong Meng, Jinan Xu, Yufeng Chen, Jinsong Su, and Jie Zhou. Towards making the most of dialogue characteristics for neural chat translation. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 67–79, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[13] Yunmeng Li, Jun Suzuki, Makoto Morishita, Kaori Abe, Ryoko Tokuhisa, Ana Brassard, and Kentaro Inui. Chat translation error detection for assisting cross-lingual communications. In **Proceedings of the 3rd Workshop on Evaluation and Comparison of NLP Systems**, pp. 88–95, Online, November 2022. Association for Computational Linguistics.

[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? In **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 2204–2213, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[15] 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.

[16] 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.

[17] Ankur Joshi, Saket Kale, Satish Chandel, and D Kumar Pal. Likert scale: Explored and explained. **British journal of applied science & technology**, Vol. 7, No. 4, p. 396, 2015.

[18] Tomoko Nemoto and David Beglar. Likert-scale questionnaires. In **JALT 2013 conference proceedings**, pp. 1–8, 2014.

[19] 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.

[20] 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, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

表 2　低品質機械翻訳モデルの学習におけるハイパーパラメータ一覧

Architecture	2-to-2 Transformer [2, 10]
Enc-Dec layers	6
Attention heads	8
Word-embedding dimension	512
Feed-forward dimension	2,048
Share all embeddings	True
Optimizer	Adam ($\beta_1 = 0.9, \beta_2 = 0.98, \epsilon = 1 \times 10^{-8}$)
Learning rate schedule	Inverse square root decay
Warmup steps	4,000
Max learning rate	0.001
Initial Learning Rate	1e-07
Dropout	0.3
Label smoothing	$\epsilon_{ls} = 0.1$
Mini-batch size	8,000 tokens
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
Implementation	`fairseq` [20]

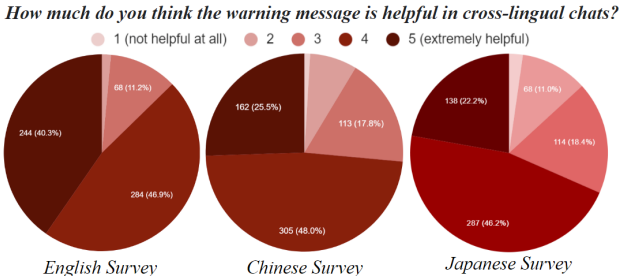

図 3　警告メッセージが対話の継続に役立ったと考えられた割合.

A　翻訳モデルの学習設定

低品質機械翻訳モデルを学習する際，まずは BPE [19] 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 [10] を学習した．表 2 にハイパーパラメータの一覧を示している．

B　結果と分析

図 3 では，3 つの言語それぞれの警告メッセージが対話の継続に役立ったと考えられた割合を示している．

様々なタイプの警告メッセージに対する収集された回答は，図 4 に要約されている．

期待される付加情報に関する結果を図 5 に示す．

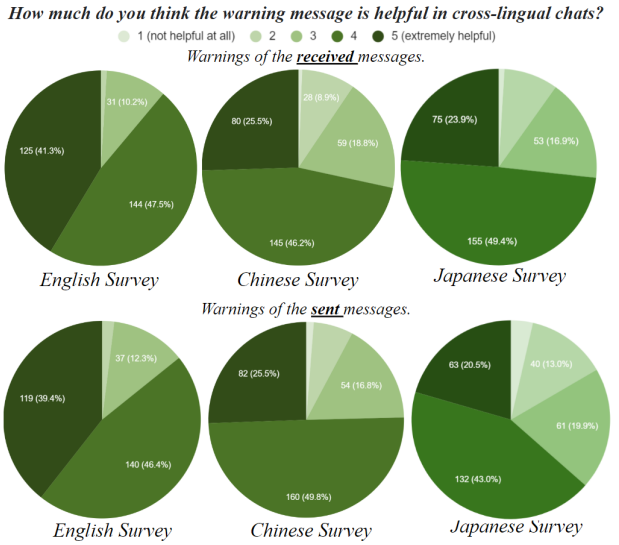

図 4　受信時または送信時における警告メッセージが対話の継続に役立ったと考えられた割合.

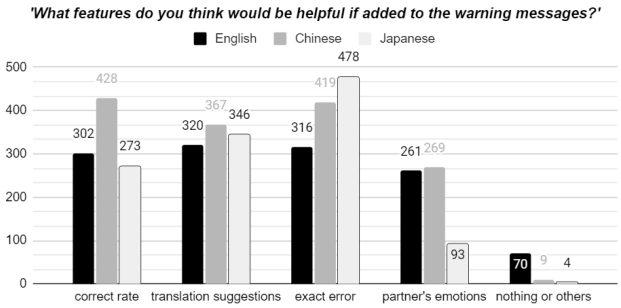

図 5　警告メッセージに期待される機能についての結果.