異言語話者の対話を仲介する音声対話翻訳

清水周一郎 [1]　Chenhui Chu[1]　Sheng Li[2]　黒橋禎夫 [1]
[1] 京都大学大学院情報学研究科 [2] 情報研究通信機構（NICT）
{sshimizu,chu,kuro}@nlp.ist.i.kyoto-u.ac.jp　sheng.li@nict.go.jp

概要

本研究では、異言語話者の対話を仲介する音声対話翻訳という新たな機械翻訳のパラダイムを提案する。日英のビジネスシーン対話コーパスに音声と話者情報を付与した SpeechBSD コーパスを構築し、ベースラインの実験を行った。音声対話翻訳においては文脈を考慮することが重要であるため、単言語文脈と二言語文脈の 2 つの設定で文脈を構成した。Whisper を用いた音声認識と mBART の fine-tuning による機械翻訳を組み合わせた cascade 音声翻訳の実験を行い、二言語文脈の有効性を示した。

1　はじめに

グローバル化に伴い、異なる言語を母語とする人々の間のコミュニケーションの重要性が高まっている。しかし、世界には言語の「壁」が存在し、コミュニケーションを妨げる大きな原因となっている。テキスト翻訳の精度は近年のニューラル機械翻訳の発展によって向上した。しかし、人間同士の意思疎通において最も重要な役割を果たす対話の分野における翻訳の研究は未だ少ない。テキストの対話であるチャットにおける対話翻訳の研究はされている [1, 2, 3] が、音声での対話翻訳の研究はされていない。

本研究では、「**音声対話翻訳**」という新しい機械翻訳のパラダイムを提案する。音声対話翻訳においては、複数の話者が異なる言語を話す多言語対話（**異言語話者の対話**）を考え、ある言語の音声を別の言語のテキストに翻訳する[1)]。応用例としては、大学のゼミをはじめとした、異なる言語を話す話者が集まるミーティングにおける利用が考えられる。また、音声対話翻訳を行う際、異なる言語の文脈を利用することで、翻訳の曖昧性を解消することを目指す（図 1）。

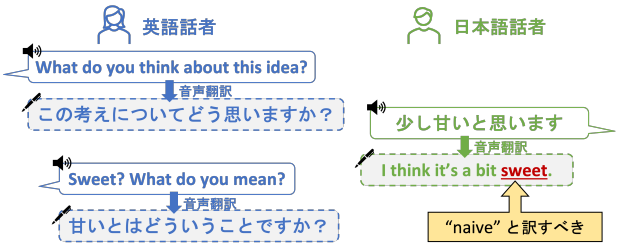


図 1　音声対話翻訳における文脈の重要性

本研究では、日英の音声対話翻訳を扱った。音声対話翻訳のデータセットは存在しないため、既存のテキスト対話コーパスであるビジネスシーン対話コーパス [4] を用いて、対応する音声をクラウドソーシングで収集して SpeechBSD データセットを構築した。また、文脈を考慮した音声対話翻訳システムとして、既存研究の手法 [5] を対話に適用した単言語で文脈を構成する手法と、二言語で文脈を構成する手法を試した。

音声翻訳の手法として、原言語の音声を目的言語のテキストに直接変換する end-to-end の手法 [6, 7] と、原言語の音声を書き起こして機械翻訳への入力とする cascade の手法 [8] があるが、本研究では cascade で実験を行った。[2)] 音声認識は、事前学習された大規模音声認識モデルである Whisper [11] を用いた。機械翻訳の実験では、文脈を利用しない場合、単言語文脈を利用する場合、二言語文脈を利用する場合の 3 つの設定で mBART [12] を用いて実験した。それらを組み合わせた cascade 音声翻訳の実験により、二言語文脈の有効性を示した。

2　関連研究

Zhang ら [5] は、音声翻訳における文脈の利用の有効性を示した。彼らは翻訳したい発話より前の発話を文脈として利用し、文脈と発話をともに翻訳するという手法を試し、その有効性を示した。文脈の

1)　音声の出力を考えることもできるが、本研究では翻訳性能に焦点を当てるため、音声合成は別のモジュールと考える。

2)　近年の研究 [9, 10] から、end-to-end と cascade の手法にほとんど性能の差は見られないことが分かっている。End-to-end ではモデルが複雑になることが多く分析が難しくなるため、本研究では casade で実験を行った。

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

利用方法を複数提案しており、本研究ではそのうちの sliding window based decoding を採用している。Zhang らの研究では 1 人の話者が講演形式で話すコーパスでの実験であったが、本研究では対話形式の状況に焦点を当てる点が異なる。

Liang ら [1] はテキストでの対話であるチャットにおける対話翻訳の研究を行い、文脈や話者情報を利用することで翻訳性能が向上することを示した。彼らは文脈を利用した翻訳タスクに加え、対話の応答予測や文脈の判定タスクなど計 5 つのタスクををマルチタスク学習し、それらが対話の翻訳に有効であることを示した。本研究はテキストではなく音声の対話に焦点を当てる点が異なるほか、新たな二言語文脈の構成方法を提案している。

3 音声対話翻訳

音声対話翻訳においては、複数の話者 S^m $(m = 1, 2, \cdots, M)$ が異なる二言語 L^n $(n = 1, 2)$ を話す異言語話者の対話を考える。**対話** $D = (U_1, \cdots, U_T)$ における**発話** U_t $(t = 1, 2, \cdots, T)$ は $U_t = (S_t^m, L_t^n, X_t)$ で表される。ここで S_t^m は発話の話者、L_t^n は発話音声の言語、X_t は発話の音声信号を表す。X_t と内容の等しい言語 L^n のテキストを Y_t^n $(n = 1, 2)$ とする。音声対話翻訳のタスクは、各発話 U_t について、音声信号 X_t から対応する翻訳 Y_t^2（原言語が L^1 の場合）または Y_t^1（原言語が L^2 の場合）を生成することである。

音声対話翻訳においては、文脈を考慮することが重要である。例として、英語話者と日本語話者が何らかの問題について対話している状況を考える（図 1）。英語話者が “Whad do you think about this idea?” と発話し、日本語話者が「少し甘いと思います」と返したとする。このとき、翻訳システムが文脈を考慮しない場合、2 つ目の発話が文脈を考慮せずに翻訳されて「甘い」を “naive” でなく “sweet” と翻訳してしまう可能性が高い。英日双方の文脈を考慮することで、この対話の文脈における「甘い」の意味が明確になり、適切な翻訳が可能になると期待される。

4 音声対話翻訳データセットの構築

音声対話翻訳のデータセットを構築するため、既存のテキスト対話コーパスであるビジネスシーン対話コーパス（以下、BSD コーパス）[4] を用いて、対応する音声と話者情報をクラウドソーシングで収集した。収集した音声と話者属性が付いた BSD コーパスを **SpeechBSD コーパス**と呼ぶこととする。

4.1 BSD コーパス

BSD コーパスは、対話における機械翻訳を推進するために人手で設計されたコーパスである。BSD コーパスはシナリオと呼ばれる単位の対話から成り、各シナリオでは 2 人以上の話者が対話している。シナリオの原文は半分が日本語、半分が英語となっており、言語によって表現が偏ることがないよう構成されている。シナリオによって含まれる文数は異なるが、平均して 30 文程度である。

4.2 構築方法

シナリオを話者ごとに分割し、クラウドソーシングにより音声を収集した。クラウドソーシングのプラットフォームとして、日本語は Yahoo!クラウドソーシング[3]、英語は Amazon Mechanical Turk[4] を用いた。音声を録音する Web ページを設計し、各プラットフォームからリンクで誘導して音声の収集を行った。録音にあたっては、静かな環境で行い、はっきりと丁寧に発音するよう指示を出した。また、録音データと共に話者の出身地（日本語の場合都道府県、英語の場合アメリカの州）及び性別の情報を集めた[5]。英語の音声の収集にあたっては、ワーカーの出身地をアメリカ合衆国に限定した。

4.3 収集した音声及び話者属性の統計

収集した音声の統計を表 1 に示す。英語音声が計 24.3 時間、日本語音声が計 30.7 時間であった。男声/女声については英語は偏りがなく、日本語ではやや男声が多かった。出身地については、日本語では人口の分布に概ねしたがっていたが、英語では分布に偏りがあった（付録図 2）。このようなデータの偏りは英語と日本語で異なるクラウドソーシングプラットフォームを利用したことに起因すると考えられる。

5 文脈を考慮した音声対話翻訳

ここでは、音声対話翻訳において文脈を考慮する手法として、単言語文脈と二言語文脈の 2 つを考える。

$M = 2$ の場合で考える。$m = n$（話者 S^i が言語 L^i $(i = 1, 2)$ で話す）としても一般性を失わない。ま

3) https://crowdsourcing.yahoo.co.jp/
4) https://www.mturk.com/
5) これらの情報の収集にあたりワーカーの同意を得ている。

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

表 1 SpeechBSD コーパスの統計

	訓練	開発	評価
シナリオ数	670	69	69
文数	20,000	2,051	2,120
英語音声（時間）	20.1	2.1	2.1
日本語音声（時間）	25.3	2.7	2.7
英語性別（男/女%）	47.2 / 52.8	50.1 / 49.9	44.4 / 55.6
日本語性別（男/女%）	68.0 / 32.0	62.3 / 37.7	69.0 / 31.0

た、簡単のため、話者は交互に発話し、話者 S^1 が対話を始めるものとする[6]。すなわち、

$$^{\forall}U \in \{U_t \mid t \bmod 2 = i\}, \quad L(U) = L^i$$

とする。

まず、S^i の時点 $t\,(t = 1, 2, \cdots, T)$ での各発話 $U_t = (S_t^i, L_t^i, X_t)$ に対し、音声認識システムを用いて書き起こし $Y_t^i\,(t = 1, 2, \cdots, T)$ を得る。音声認識における文脈の必要性は機械翻訳と比べて小さいと仮定し、ここでは文脈を考慮しない。

1 つ目の単言語文脈を利用する手法は、書き起こしを機械翻訳したテキスト[7]を用いて単言語で文脈を構成し、その文脈を用いて翻訳したい発話を翻訳する手法である。発話 U_τ の言語 L^i についての単言語文脈のテキストは

$$\boldsymbol{Y}_{<\tau}^i = \{Y_t^i \mid t < \tau\}, \quad i = 1, 2$$

で表せる。この場合、機械翻訳システムは単言語対のものが 2 つでも良いし、二言語対に対応した 1 つの翻訳システムでも良い。本研究では単言語対の機械翻訳システム 2 つを用いた。L^1 を原言語、L^2 を目的言語とする機械翻訳モデルは、以下の対数尤度を最大化するように学習する。

$$\mathcal{L}^{1\to 2} = \sum_t \log \mathrm{P}(Y_t^2, \boldsymbol{Y}_{<t}^2 \mid Y_t^1, \boldsymbol{Y}_{<t}^1)$$

文脈及び翻訳対象の発話は発話順 t に沿って与える。出力は Y_t^2 と $\boldsymbol{Y}_{<t}^2$ が共に出てくるため、後処理により Y_t^2 のみを取り出す。L^2 を原言語、L^1 を目的言語とする場合も同様である。

2 つ目の二言語文脈を利用する手法は、文脈を二言語の書き起こし[8]から構成する手法である。発話 U_τ の二言語文脈のテキストは $\boldsymbol{Y}_{<\tau} = \tilde{\boldsymbol{Y}}_{<\tau}^1 \cup \tilde{\boldsymbol{Y}}_{<\tau}^2$ で表せる。ここで

$$\tilde{\boldsymbol{Y}}_{<\tau}^i = \{Y_t^i \mid t < \tau \wedge t \bmod 2 = i\}, \quad i = 1, 2$$

である。この場合、機械翻訳システムは二言語対に対応したものである必要がある。$\boldsymbol{Y}_{<\tau}$ の翻訳を $\overline{\boldsymbol{Y}_{<\tau}} = \overline{\tilde{\boldsymbol{Y}}_{<\tau}^1} \cup \overline{\tilde{\boldsymbol{Y}}_{<\tau}^2}$ とする。ここで

$$\overline{\tilde{\boldsymbol{Y}}_{<\tau}^i} = \{Y_t^j \mid t < \tau \wedge t \bmod 2 = i\}, \quad (i, j) = (1, 2), (2, 1)$$

である。二言語文脈を扱う機械翻訳モデルは、以下の対数尤度を最大化するように学習する。

$$\mathcal{L} = \sum_t \log \mathrm{P}(\overline{Y_t}, \overline{\boldsymbol{Y}_{<\tau}} \mid Y_t, \boldsymbol{Y}_{<\tau})$$

ここで、$\overline{Y_t}$ は $L(U_t) = L^1$ のとき Y_t^2、$L(U_t) = L^2$ のとき Y_t^1 とする。文脈及び翻訳対象の発話は発話順 t に沿って与え、出力は後処理により $\overline{Y_t}$ のみを取り出す。

なお、文脈 $\boldsymbol{U}_{<\tau} = \{U_t \mid t < \tau\}$ を考える際、モデルの入力長には制限があるため、文脈幅 c を考慮する。発話 U_τ の文脈幅が c の文脈は $\boldsymbol{U}_{<\tau} = \{U_t \mid t = \tau - 1, \cdots, \tau - c \wedge t > 0\}$ で表せる。

6 実験

6.1 音声認識

本研究は日英の翻訳を対象とするため、日本語と英語の音声認識器が必要となる。本研究では、多言語音声認識及び英語を目的言語とする多言語音声翻訳のマルチタスクモデルである Whisper [11] を用いた。Whisper は log-Mel spectrogram 特徴量を入力に用いた Transformer [13] ベースのモデルで、計 680,000 時間の音声データを用いて教師あり学習されており、多様なドメインのデータで訓練することによって頑健性の高い性能を達成している。encoder 12 層、decoder 12 層の medium を fine-tuning 等を行わずに用いた。語彙は公開されている語彙サイズ 50,257 の byte-level BPE モデルを用いた。SpeechBSD コーパスの評価データで音声認識の予測を行い、日本語は文字誤り率 (CER)、英語は単語誤り率 (WER) で評価した。結果は日本語の CER が 13.2%、英語の WER が 8.3% であった。

6.2 機械翻訳

本研究では翻訳の際に二言語の文脈を考慮するため、二言語対に対応した機械翻訳モデルを用いる必

6) 本研究で用いるデータセットでは、同じ話者による連続した複数の発話は別の発話として扱うため、話者が交互に発話するわけではない。また、話者が 3 者以上の場合、対話内での登場順に番号を振り、番号の偶奇が一致する話者は同じ言語を話すものとして扱う。

7) 推論時には書き起こしを用いるが、訓練時にはそれに対応する正解のテキストを用いる。

8) 同様に推論時のみで、訓練時にはそれに対応する正解のテキストを用いる。

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

要がある。本研究では mBART [12] の 25 言語版（日本語を含む）の large（各 12 層の encoder と decoder から成る Transformer）モデルを用いて、公開されている語彙（語彙サイズ 250,001）を用いて BSD コーパスで fine-tuning を行った。実装には Fairseq [14] を用いた。設定として文脈を利用しない場合、単言語文脈を利用する場合、二言語文脈を利用する場合の 3 つを試した。

6.2.1 実験設定

文脈を利用しない場合 対話の各発話を個別のものとして扱い、通常の機械翻訳と同じ方法で mBART の fine-tuning を行った。モデルは日英・英日で個別のものとした。前処理として、日本語・英語の各文に mBART の sentencepiece model を適用してサブワード分割を行った。訓練時のハイパーパラメータは基本的に mBART [12] の論文と同じものを用いたが、開発データの loss に基づく patience 10 の early stopping を採用した。Checkpoint を epoch ごとに保存し、翻訳生成時のモデルには最後の 10 checkpoint を平均したものを用いた。評価には SacreBLEU [15] を用いた[9]。

単言語文脈を利用する場合 各シナリオで 5 節の方法で文脈幅 $c = 2$[10] の単言語文脈を構成し、mBART の fine-tuning を行った。モデルは日英・英日で個別のものとした。文脈及び翻訳する発話同士は end of sentence トークン</s>で繋げた。そのほかのパラメータは文脈を利用しない場合と同じである。

二言語文脈を利用する場合 各シナリオで 5 節の方法で文脈幅 $c = 2$ の二言語文脈を構成し、mBART の fine-tuning を行った。文脈及び翻訳する発話同士は end of sentence トークン</s>で繋げた。mBART では入出力の最後のトークンとして ja_XX や en_XX といった言語タグを用いて翻訳の言語対を指定する必要がある。二言語文脈を使用する場合には原言語や目的言語といった考え方ができないため、入力側に ja_XX、出力側に en_XX をつける場合とその逆の場合を試し、BLEU スコアはその平均をとった（両者に顕著な差は見られなかった。）。また、評価の際には文脈を利用しない場合と同じ出力形式となるように戻して評価した。そのほかのパラメータは文脈を利用しない場合と同じである。

9） 英日: nrefs:1|case:mixed|eff:no|tok:ja-mecab-0.996-IPA|smooth:exp|version:2.0.0
日英: nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.0.0

10） Zhang らの研究で性能の高かった値を採用した。

表 2 機械翻訳 (MT) 及び cascade 音声翻訳 (Cascade ST) の BSD コーパス評価データの BLEU スコア

		英日	日英
MT	文脈なし	15.9	18.2
	文脈あり（単言語）	**16.8**	19.1
	文脈あり（二言語）	16.7	**19.7**
Cascade ST	文脈なし	15.2	15.4
	文脈あり（単言語）	15.9	15.8
	文脈あり（二言語）	**16.1**	**17.1**

6.2.2 結果

表 2 に文脈を用いなかった場合、単言語文脈を用いた場合、二言語文脈を用いた場合の比較を示す。文脈を用いなかった場合と比べて、単言語文脈の利用により英日・日英とも 0.9 ポイントの BLEU スコアの改善が見られた。単言語文脈と二言語文脈を使用した場合の結果に顕著な差は見られなかった。

6.3 Cascade 音声翻訳

音声認識の出力を機械翻訳への入力とすることで cascade 音声翻訳の実験を行った。手法は 5 節で記述した通りである。音声認識には Whisper の出力結果を用い、文脈を利用しない場合、単言語文脈を利用する場合、二言語文脈を利用する場合のそれぞれで実験した。

結果を表 2 に示す。音声翻訳においても、文脈を用いなかった場合と比べて、単言語文脈を利用することで 0.4 - 0.7 ポイントの BLEU スコアの改善が見られた。また、二言語文脈を利用した場合はさらに 0.2 - 1.3 ポイントの改善が見られた。

7 おわりに

本研究では、音声対話翻訳という新たな研究の枠組みを提案し、クラウドソーシングを用いた音声収集により SpeechBSD コーパスを構築した。文脈の利用方法として単言語文脈を利用する手法と二言語文脈を利用する手法を試し、実験によって二言語文脈の有効性を示した。今後の展望として、end-to-end の音声翻訳モデルの実験や、話者の属性情報を用いた音声翻訳を行う予定である。

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

謝辞

本研究は JSPS 科研費 JP19K20343 及びサムスン SDS 株式会社の助成を受けたものである。

参考文献

[1] 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.

[2] 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.

[3] Yunlong Liang, Fandong Meng, Jinan Xu, Yufeng Chen, and Jie Zhou. Scheduled multi-task learning for neural chat translation. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 4375–4388, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[4] 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.

[5] Biao Zhang, Ivan Titov, Barry Haddow, and Rico Sennrich. Beyond sentence-level end-to-end speech translation: Context helps. 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. 2566–2578, Online, August 2021. Association for Computational Linguistics.

[6] Long Duong, Antonios Anastasopoulos, David Chiang, Steven Bird, and Trevor Cohn. An Attentional Model for Speech Translation Without Transcription. In **Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, 2016.

[7] Alexandre Berard, Olivier Pietquin, Christophe Servan, and Laurent Besacier. Listen and Translate: A Proof of Concept for End-to-End Speech-to-Text Translation. In **NIPS Workshop on End-to-end Learning for Speech and Audio Processing**, 2016.

[8] F. W. M. Stentiford and M. G. Steer. Machine translation of speech. **British Telecom technology journal**, 1988.

[9] Luisa Bentivogli, Mauro Cettolo, Marco Gaido, Alina Karakanta, Alberto Martinelli, Matteo Negri, and Marco Turchi. Cascade versus direct speech translation: Do the differences still make a difference? 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. 2873–2887, Online, August 2021. Association for Computational Linguistics.

[10] Viet Anh Khoa Tran, David Thulke, Yingbo Gao, Christian Herold, and Hermann Ney. Does Joint Training Really Help Cascaded Speech Translation?, 2022.

[11] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Chiristine McLeavey, and Ilya Sutskever. Robust Speech Recognition via Large-Scale Weak Supervision, 2022.

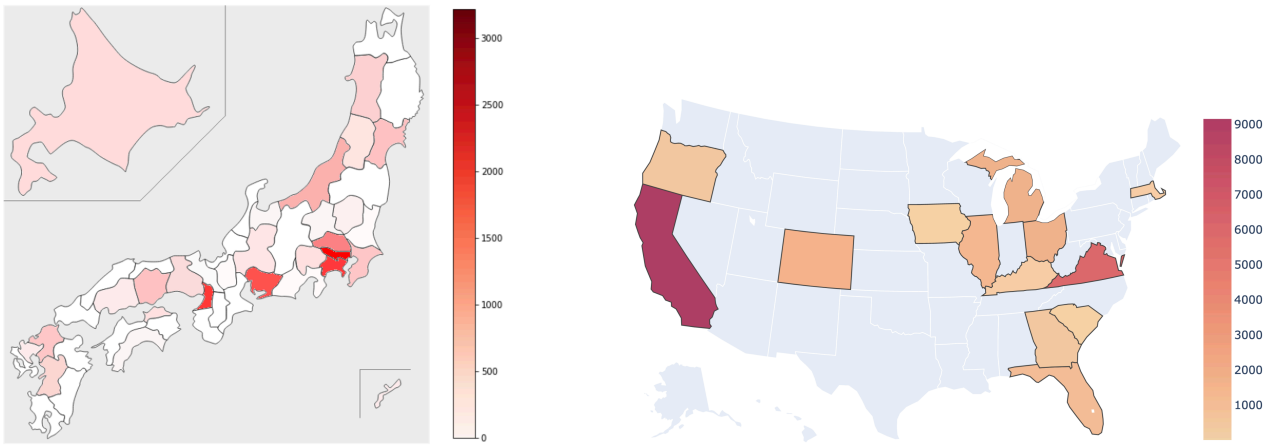
[12] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. **Transactions of the Association for Computational Linguistics**, Vol. 8, pp. 726–742, 2020.

[13] 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 I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, **Advances in Neural Information Processing Systems**, Vol. 30. Curran Associates, Inc., 2017.

[14] 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.

[15] Matt Post. A call for clarity in reporting BLEU scores. In **Proceedings of the Third Conference on Machine Translation: Research Papers**, pp. 186–191, Brussels, Belgium, October 2018. Association for Computational Linguistics.

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).



(a) 日本語　　(b) 英語

図 2　収集した音声の話者の出身地の発話数による分布

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).