非流暢性タグを用いた目的言語テキストによる自由発話の音声翻訳

胡 尤佳　須藤 克仁　中村 哲
奈良先端科学技術大学院大学
{ko.yuka.kp2,sudoh,s-nakamura}@is.naist.jp

概要

自由発話はフィラーや言い淀みなどの非流暢性を含むことがあるため，発話を目的言語のテキストや音声に翻訳する音声翻訳でもそれらの現象への対応が重要である．特に，非流暢性を含む音声から流暢なテキストへの翻訳は，翻訳時に非流暢性を除去する必要があり，非流暢性を含むテキストへの翻訳よりも難しい．本研究では，非流暢性タグを含む目的言語テキストで学習したモデルを用い，流暢なテキストへの翻訳を学習する手法を提案する．本研究で用いるタグは，多様な非流暢性を一つのタグで表し，非流暢性事象の位置情報を含む．実験から，提案手法による非流暢性を含む音声から流暢なテキストへの翻訳性能向上の効果が示された．

1　はじめに

人間の自然発話においてはフィラーや言い淀みといった非流暢性がしばしば含まれ，それらの現象を取り扱うことのできる音声認識 (ASR)，音声翻訳 (ST) が実用上求められている．通常，発話に非流暢性が含まれていると，ASR の出力テキストも非流暢性を含み，機械翻訳 (MT) での翻訳が難しい．その結果，ASR と MT を繋ぎ合わせたパイプライン型音声翻訳での最終的な性能低下につながる．それを防ぐため，非流暢性検出器により ASR の出力から検出された非流暢性を取り除いてから MT に入力する手法が主流である [1]．近年の音声翻訳では，非流暢性を含む音声と非流暢性を含む目的言語テキスト [2]，もしくは流暢な目的言語テキスト [3] により構成される音声翻訳データにより，明示的な非流暢性検出器を用いずに End-to-End 音声翻訳モデルを学習する手法が主流になりつつある [3]．

本研究では，可読性が高い出力をより低遅延で得られる音声翻訳を想定し，非流暢性を含む音声から流暢なテキストへの翻訳の性能向上を目指す．非流暢性を含む音声翻訳での課題として，データが少量である上，特に非流暢な音声から流暢なテキストへの翻訳では，翻訳時に非流暢性を除去し，可読性が高く要約された出力を得る必要があるため，非流暢性を含むテキストへの翻訳と比較して難易度が高い．このような限られたデータを最大限に利用し，非流暢性を含むテキストへの翻訳モデルを作成し，そのモデルを用いて流暢なテキストへの翻訳を学習することで，流暢なテキストへの翻訳性能の向上が期待できる．しかし，非流暢性を含むテキストで学習したモデルを直接利用して流暢なテキストへの翻訳を学習した場合，非流暢性の表記が複数ある上，どの部分が非流暢性であるかという情報を含まないため，曖昧性が高く性能向上の妨げとなり得る．

Horii らは日本語音声認識において非流暢性をグループに分類し，それぞれの非流暢性を意味する非流暢性タグを用いて該当部分を置き換え学習することにより，音声認識の性能を向上させた [4]．この方法では，多様な非流暢性をタグにまとめることによりタスクを簡略化している．また，出力がタグ付きのテキストになるため非流暢性の位置の特定が可能であり，タグを取り除くことで流暢なテキストに近いテキストを生成することも可能である．

本研究は，非流暢性タグを音声翻訳データの目的言語テキストで取り入れ，タグを含む学習データを用いてモデルを学習する．その後，このデータで学習したモデルを用いて流暢なテキストへの翻訳を学習する．Horii らの研究では，音声認識の性能向上を目的とし，原言語テキストに含まれる，あらかじめ人手で非流暢性とアノテーションされた部分をタグで置き換えている．本研究では，音声翻訳の性能向上を目的とし，目的言語テキストに含まれる非流暢性をタグで置き換えて学習し，その際に，目的言語テキストがアノテーションされた非流暢性の情報

を持たないことを想定し，非流暢性検出器で非流暢性と検出された部分をタグで置き換える手法を取り入れる．実験により，非流暢性タグを用いたデータを用いて学習したモデルにより，非流暢性を含む音声から流暢なテキストへの翻訳での性能向上が期待できることが示された．

2　音声翻訳

$\mathbf{X} = (x_1, ..., x_T)$ を原言語の入力音声に対する音響特徴量の系列，$\mathbf{S} = (s_1, ..., s_M)$ を原言語テキストのトークン系列，$\mathbf{T} = (t_1, ..., t_N)$ を目的言語テキストのトークン系列とする．v を語彙集合 V の元とすると，t_i の事後確率は以下の式で表される．

$$P_{\mathrm{ST}}(t_i = v) = p(v|\mathbf{X}, t_{<i}). \tag{1}$$

ST の学習時の損失関数 $\mathcal{L}_{\mathrm{ST}}$ は，Cross-entropy loss を用いて以下の式で表される

$$\mathcal{L}_{\mathrm{ST}} = -\sum_{i=1}^{N}\sum_{v\in V}\delta(v, t_i)\log P_{\mathrm{ST}}(t_i = v). \tag{2}$$

$\delta(v, t_i)$ は，$v = t_i$ のとき 1，そうでなければ 0 となる関数である．

3　非流暢性タグを用いたデータの生成と学習

提案手法で用いる非流暢性タグを用いたデータの生成過程を図 1 に示す．提案手法では，目的言語テキストのどの単語が非流暢性かという情報が明示的に付与されていない音声翻訳データを用いることを前提とし，Lou らによる，BERT を用いて事前学習された非流暢性検出器 [5, 6] を用いて非流暢性タグを付与する．この非流暢性検出器は，構文解析と非流暢性検出のタスクをマルチタスク学習により同時に学習することで得られたモデルであり，本研究では，本検出器により得られたタグ付きの音声翻訳データを学習データとして用いる．モデルの学習で利用するデータは，非流暢性検出器から得られた非流暢性の情報をもとに，以下二種類のものを作成する．

- $D2D_{Tag}$: 目的言語テキストの非流暢性をタグで置き換えたデータ
- $D2D_{NoTag}$: 目的言語テキストの非流暢性を取り除いたデータ

また，学習済みモデルに対して Fine-tuning をするためのデータとして，非流暢性が検出された文のみで構成されたデータ $\Delta D2D_{Tag}, \Delta D2D_{NoTag}$ も作成する．

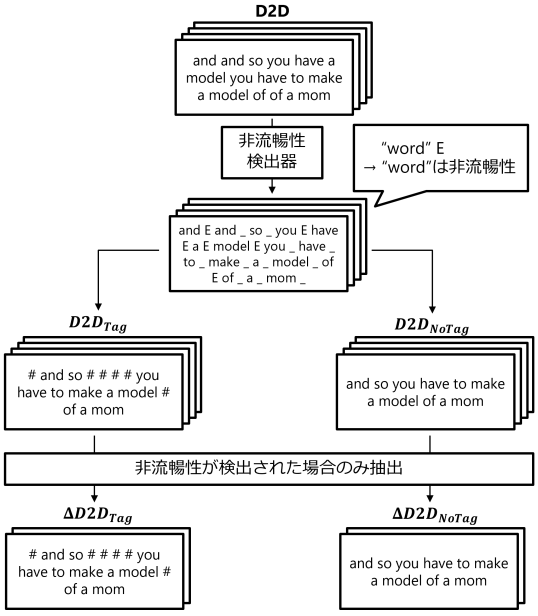

図 1　非流暢性検出器を用いた非流暢性タグが付与された，もしくは除かれたデータ作成の概略図．

4　実験

非流暢性を含む音声から流暢なテキストへの翻訳で，提案手法である非流暢性タグを用いたデータで学習したモデルによる性能向上がみられるかを検証するため，以下の実験を行った．

4.1　データセット

本研究では音声翻訳データとして，Fisher Spanish Corpus (Spanish-English) を用いた [2, 3]．本研究における Fisher Spanish Corpus は，以下からなるデータで構成されたものを用いた．

- $D2D$: 非流暢性を含む音声から非流暢性を含む目的言語テキスト
- $D2F$: 非流暢性を含む音声から流暢な目的言語テキスト

モデルは Fairseq [7] を用いて実装し，Transformer [8] をベースに作成した．音響特徴量は 80 次元のメルフィルタバンク特徴量を用い，学習データでは SpecAugument [9] によるデータ拡張手法を用いた．Tokenizer は SentencePiece [10] を用い，最大語彙数は 8000 とした．提案手法においては，非流暢性を扱うために新たに定義した < # > を非流暢性ラベルとして定義し，それを含めた語彙数 8000 の辞書を作成した．学習データからは，文字数が 400 より大きく，フレーム数が 3000 より大きいペアは取り

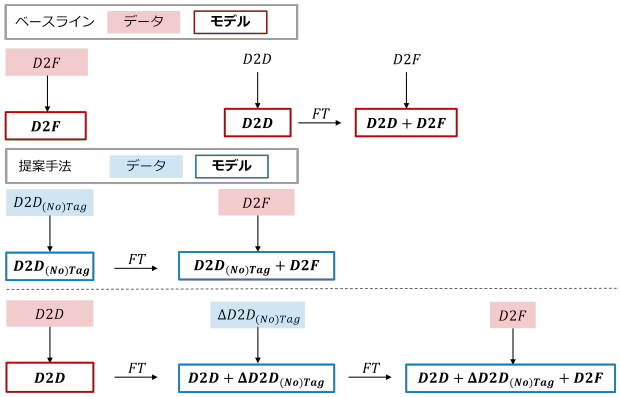

図 2　ベースラインと提案手法の学習過程の概略図.

表 1　学習に用いた Fisher Spanish Corpus の内訳．() 内はオリジナルでのデータサイズ.

	非流暢性を含むテキスト	流暢なテキスト
Train	138770 (138819)	138718 (138720)
Dev	3977 (3979)	3977 (3977)
Test	3641	3641

除いた．学習データの内訳を表 1 に示す.

4.2　End-to-End 音声翻訳モデル

本研究では，ベースラインと提案手法として以下のモデルを用いた．ベースラインと提案手法モデルの学習過程を図 2 に示す.

4.2.1　ベースラインモデル

本実験では，以下モデルをベースラインとした.

- *D2D*: 非流暢性を含む音声から非流暢性を含む目的言語テキストへの翻訳を学習したモデル
- *D2F*: 非流暢性を含む音声から流暢な目的言語テキストへの翻訳を学習したモデル
- $D2D + D2F$: *D2D* をもとに流暢な目的言語テキストへの翻訳を Fine-tuning で学習したモデル

4.2.2　提案手法モデル

提案手法では，図 1 で作成したデータを用いてモデルを作成した．非流暢性を含む目的言語テキストから非流暢性検出器により検出されたタグの内訳を表 2 に示す．$D2D_{Tag}, D2D_{NoTag}$ は，非流暢性を含む目的言語テキストと同じ数のデータとなっており，学習されたモデルは，タグ情報を用いていない非流暢性を含むテキストをもとにしたベースラインの *D2D* モデルと比較できる．$D2D_{Tag}, D2D_{NoTag}$ を用いたモデルとして，$D2D_{Tag}, D2D_{Tag}{+}D2F, D2D_{NoTag}, D2D_{NoTag}{+}D2F$

表 2　目的言語テキストに含まれる非流暢性タグの割合.

	タグを含む文の割合	タグの割合
Train	29527 / 138770 (21.3%)	67592 / 1439553 (4.7%)
Dev	848 / 3977 (21.3%)	1834 / 39966 (4.7%)
Test	731 / 3641 (20.1%)	1760 / 39538 (4.5%)

を作成した.

また，$\Delta D2D_{Tag}, \Delta D2D_{NoTag}$ データを用いて，*D2D* モデルで Fine-tuning したモデルを作成した．$\Delta D2D_{Tag}, \Delta D2D_{NoTag}$ データを用いたモデルとして，$D2D + \Delta D2D_{Tag}, D2D + \Delta D2D_{Tag} + D2F, D2D + \Delta D2D_{NoTag}, D2D + \Delta D2D_{NoTag} + D2F$ を作成した.

4.3　評価

本研究では，以下の評価データを用いて SacreBLEU [11] により評価した.

- *D2F test*：非流暢性を含む音声から流暢な目的言語テキスト
- $D2D_{NoTag}$ *test*：非流暢性を含む音声から非流暢性タグを除いた目的言語テキスト
- $D2D_{Tag}$ *test*：非流暢性を含む音声から非流暢性タグを含む目的言語テキスト

$D2F$ *test*, $D2D_{NoTag}$ *test* を用いる際には，予測結果の非流暢性タグは除去した.

5　実験結果と議論

ベースラインと提案手法の BLEU の結果を表 4 に示す．実験結果から，提案手法による非流暢性タグの情報を用いたモデルを流暢なテキストで学習したモデルで，$D2D_{Tag} + D2F, D2D_{NoTag} + D2F, D2D + \Delta D2F_{NoTag} + D2F$ が，いずれもベースラインと比較して *D2F test* での BLEU の向上が見られることがわかった．この結果から，非流暢性タグの情報を取り入れた非流暢性を含むテキストへの翻訳を学習した場合の方が，流暢な目的言語テキストへの翻訳において有効だと分かった.

一方で，流暢なテキストで Fine-tuning をしない場合の，非流暢性タグの情報を用いたモデル $D2D_{Tag}, D2D + \Delta D2D_{Tag}, D2D_{NoTag}, D2D + \Delta D2D_{NoTag}$ では，*D2F test* での評価値がベースラインである $D2D + D2F$ の評価値を上回らなかった．このことから，非流暢性タグを付与した，もしくは除いて学習したモデルは，非流暢性部分が予測されるよう，もしくは除かれるよう学習されるものの，

表 3 *D2F test* で評価した際のベースラインと提案手法の例文.

ベースライン	
D2D	yes sure that's that's that's that's it that's it's an advantage that's true that there's less cold
D2D + *D2F*	yes that's that's definitely it's an advantage that there's less cold
提案手法	
$D2D_{Tag}$	yes sure < # >< # >< # > this is definitely it's an advantage that there's less cold
$D2D_{Tag}$ + *D2F*	yes sure **that's** definitely it's an advantage right that there's less rules
D2F test	that is definitely an advantage

表 4 ベースラインと提案手法の BLEU の結果 (下線はベースラインより評価値が向上したもの).

	test	
	D2F	$D2D_{NoTag}$
ベースライン		
D2F†	9.1	13.7
D2D†	10.7	22.8
D2D + *D2F*	11.8	22.0
提案手法		
$D2D_{Tag}$†	9.3	21.1
$D2D_{Tag}$ + *D2F*	12.2	22.4
$D2D_{NoTag}$†	9.8	23.4
$D2D_{NoTag}$ + *D2F*	12.1	22.1
D2D + $\Delta D2D_{Tag}$	5.1	16.6
D2D + $\Delta D2D_{Tag}$ + *D2F*	11.5	20.7
D2D + $\Delta D2D_{NoTag}$	10.4	25.1
D2D + $\Delta D2D_{NoTag}$ + *D2F*	12.1	21.8

流暢な翻訳に最適化されていないことが分かった.

$D2D_{NoTag}$ *test* での評価では，提案手法である $D2D_{NoTag}$ と *D2D* + $\Delta D2D_{NoTag}$ で，ベースラインである *D2D* と比較して性能向上が見られており，非流暢性を除去したテキストへの翻訳をする上では，非流暢性を除去してから学習したモデルが最も最適化されていることが分かった．その一方で，$D2D_{Tag}$ と *D2D* + $\Delta D2D_{Tag}$ のようなタグを含むテキストで学習したモデルで出力に含まれるタグを除去し $D2D_{NoTag}$ *test* での評価もしたが，ベースラインと比べて性能向上は見られなかった.

タグを含むデータで学習したモデルである $D2D_{Tag}$と *D2D* + $\Delta D2D_{Tag}$ を $D2D_{Tag}$ *test* で評価した結果と，出力に含まれる非流暢性タグの割合を表 5 に示す．$D2D_{Tag}$ *test* での評価値は，ベースラインの *D2D* を含めほとんどのモデルでの

† これらのモデルは，非流暢性を含む音声と非流暢性を含む原言語テキストのペアで学習した音声認識モデルにより，Encoder のパラメータを初期化している.

表 5 $D2D_{Tag}$, *D2D* + $\Delta D2D_{Tag}$ モデルを $D2D_{Tag}$ *test* で評価した際の BLEU と出力に含まれる非流暢性タグの割合.

	$D2D_{Tag}$ *test*	タグの割合
$D2D_{Tag}$ *test*	-	1760 / 39538 (4.5%)
$D2D_{Tag}$	16.9	8926 / 41543 (21.5%)
D2D + $\Delta D2D_{Tag}$	12.4	16877 / 44366 (38.0%)

$D2D_{NoTag}$ *test* での評価値と比較し低い値となっている．出力に含まれる非流暢性タグの割合をみると，参照文である $D2D_{Tag}$ *test* では 4.5% のみタグが含まれているのに対し，*D2D* + $\Delta D2D_{Tag}$ では 38.0%，$D2D_{Tag}$ では 21.5% とより多くの非流暢性タグが含まれており，参照文中に含まれるタグと出力文に含まれるタグの数の差が大きいことが，評価値の低下の要因だと考えられる.

タグを付与したデータで学習したモデルは，このようなタグの過剰生成の傾向があるものの，$D2D_{Tag}$ + *D2F* では性能の向上が見られた．ベースラインと提案手法の一つである $D2D_{Tag}$,$D2D_{Tag}$ + *D2F* での出力文の例を表 3 に示す．$D2D_{Tag}$ の出力例では，非流暢性である連続した“that's”の部分に，適切に非流暢性タグが付与されていることが分かる．また，その後流暢なテキストで Fine-tuning した $D2D_{Tag}$ + *D2F* では，ベースラインである *D2D* + *D2F* で見られた“that's”の繰り返しを防げており，提案手法の有効性が確認できた.

6 まとめと今後の課題

本研究では，目的言語テキストに対して非流暢性タグを付与したデータを用い，非流暢性を含む音声から流暢な目的言語テキストの翻訳を学習した．実験により，流暢な目的言語テキストへの翻訳での性能向上が期待できることが分かった．今後の課題として，出力文に含まれる非流暢性タグの数をコントロールできるような学習と，出力位置の予測精度の向上を考えている.

謝辞

本研究の一部は JSPS 科研費 JP21H05054 と JST SPRING プログラム JPMJSP2140 の助成を受けたものである.

参考文献

[1] Eunah Cho, Thanh-Le Ha, and Alex Waibel. Crf-based disfluency detection using semantic features for german to english spoken language translation. In **Proceedings of the 10th International Workshop on Spoken Language Translation: Papers, Heidelberg, Germany, December 5-6, 2013**, 2013.

[2] Matt Post, Gaurav Kumar, Adam Lopez, Damianos Karakos, Chris Callison-Burch, and Sanjeev Khudanpur. Improved speech-to-text translation with the fisher and callhome Spanish-English speech translation corpus. In **Proceedings of the 10th International Workshop on Spoken Language Translation: Papers**, Heidelberg, Germany, December 5-6 2013.

[3] Elizabeth Salesky, Matthias Sperber, and Alexander Waibel. Fluent translations from disfluent speech in end-to-end speech translation. 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. 2786–2792, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[4] Koharu Horii, Meiko Fukuda, Kengo Ohta, Ryota Nishimura, Atsunori Ogawa, and Norihide Kitaoka. End-to-end spontaneous speech recognition using disfluency labeling. In Hanseok Ko and John H. L. Hansen, editors, **Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022**, pp. 4108–4112. ISCA, 2022.

[5] Paria Jamshid Lou, Yufei Wang, and Mark Johnson. Neural constituency parsing of speech transcripts. 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. 2756–2765, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[6] Paria Jamshid Lou and Mark Johnson. Improving disfluency detection by self-training a self-attentive model. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 3754–3763, Online, July 2020. Association for Computational Linguistics.

[7] 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 Waleed Ammar, Annie Louis, and Nasrin Mostafazadeh, editors, **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Demonstrations**, pp. 48–53. Association for Computational Linguistics, 2019.

[8] 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**, pp. 5998–6008, 2017.

[9] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. Specaugment: A simple data augmentation method for automatic speech recognition. In Gernot Kubin and Zdravko Kacic, editors, **Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019**, pp. 2613–2617. ISCA, 2019.

[10] Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations**, pp. 66–71, 2018.

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