漸進的係り受け解析と残存文長推定に基づく講演文への逐次的な改行挿入

高橋 晨成[1]　大野 誠寛[1]　松原 茂樹[2]
[1] 東京電機大学 大学院未来科学研究科　[2] 名古屋大学 情報基盤センター
23fmi15@ms.dendai.ac.jp　ohno@mail.dendai.ac.jp
matsubara.shigeki.z8@f.mail.nagoya-u.ac.jp

概要

講演を対象とした字幕生成システムにおいて，講演文特有の長い文が複数行にまたがって表示される際に，適切な位置に改行を挿入し，読みやすい字幕を生成する必要がある．これまでにも，改行挿入手法がいくつか提案されているが，特に逐次的な改行挿入手法には精度向上の余地が残されている．

そこで本稿では，読みやすい字幕を生成するための要素技術として，漸進的係り受け解析と残存文長推定に基づく逐次的な改行挿入手法を提案する．本手法の特徴は，各文節が入力されるごとに改行を挿入するか否かを BERT により判定する際に，同時に残存文長の推定を行いつつ，漸進的係り受け解析から得られる構文情報を用いる点にある．

1　はじめに

聴覚障碍者や高齢者，外国人らの講演音声の聞き取りや理解を支援するために，字幕を生成することが望まれている．講演では一文が長くなる傾向があり，複数行にまたがって表示される場合，適切な位置に改行を挿入し，読みやすい字幕を生成する必要がある．これまでにも，改行挿入手法がいくつか提案されている [1, 2, 3, 4]．従来研究 [1] は，1 文全体に対して適切な改行位置を求める手法を提案している．しかし，1 文全体の発話が終わるまで改行位置の推定を行えないため，リアルタイムでの字幕生成には必ずしも適さない．それに対し従来研究 [2] は，文節が入力されるたびに，その直前の文節との間に改行を挿入するか否かを機械学習（最大エントロピー法）を用いて，逐次的に推定する手法を提案している．しかしながら，1 文全体の係り受け構造や未入力部分の情報を使えないという制約があり，1 文全体に対して改行位置の推定を行う従来手法 [1] と比べて，その精度は低く，改善の余地がある．

逐次的な処理においても詳細な係り受け情報を使用する試みとして，従来研究 [3] では，漸進的係り受け解析手法を提案するとともに，それを応用した逐次的な改行挿入手法を提案している．この漸進的係り受け解析手法 [3] では，文節が入力されるごとに，既入力文節の係り先が他の既入力文節のいずれであるか，あるいは，未入力文節であるかを同定し出力することができ，この解析結果から得られる係り受け情報に関する素性を用いた逐次的な改行挿入手法となっている．さらに，未入力部分の情報を補う試みとして，従来研究 [4] は，文の残りの長さである残存文長が短いほど改行の必要性は低下すると考え，BERT（Bidirectional Encoder Representations from Transformers）[5] を用いて，残存文長を推定しつつ，改行を挿入するか否かを逐次的に判定する手法を提案している．両手法 [3, 4] とも精度がそれぞれ向上しており，逐次的な改行挿入において，漸進的係り受け解析結果と残存文長をともに考慮することにより，更に精度が向上する可能性がある．

そこで本稿では，読みやすい字幕をリアルタイムに生成するための要素技術として，漸進的係り受け解析と残存文長推定に基づく逐次的な改行挿入手法を提案する．提案手法では，各文節が入力されるごとに，改行挿入判定と残存文長推定を BERT を用いて同時実行し，その際の改行挿入判定において，漸進的係り受け解析から得られる構文情報を用いることにより，逐次的な改行挿入における精度向上を試みる．なお，従来の改行挿入手法 [1, 2, 3, 4] では，音響情報としてポーズの長さのみが使われている．様々な音声言語処理（例えば [6]）において一般的に使われる音響情報として，ポーズ以外にも，声の高さや強さ，発話速度があり，有効であることが知られている．提案手法では，これらの音響情報も新た

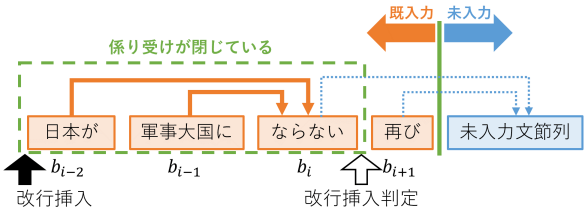

図 1　素性 f_{12} を判定する際に用いる係り受け構造

に用いる．

2　従来の逐次的改行挿入手法

2.1　基本素性に基づく逐次的改行挿入

従来手法 [2] では形態素情報，文節まとめ上げ，節境界解析，係り受け解析が施された文節列を入力とし，$i+1$ 番目の文節 b_{i+1} が入力されるたびに，その直前の文節との境界，すなわち b_i の直後に改行を挿入するか否かの判定を最大エントロピー法（ME）を用いて逐次的に行う．ME の素性には，b_i の主辞，語形などの形態素情報 4 種 $f_1 \sim f_4$，節境界情報 2 種 $f_5 \sim f_6$，係り受け情報 3 種 $f_7 \sim f_9$，行頭からの文字数 1 種 f_{10}，ポーズ情報 1 種 f_{11} の合計 11 種の基本素性を使用している．ここで，係り受け情報 3 種 $f_7 \sim f_9$ は，隣接文節に係るか否かの解析結果から得られる素性であり，文節列間の係り受け構造を利用しているわけではないことに注意されたい．なお，ディスプレイの大きさを考慮して 1 行の最長文字数を 20 字と設定し，b_i の直後に改行を挿入しなければ最長文字数を超える場合には，ME の判定結果に関わらず強制的に改行を挿入する．

2.2　漸進的係り受け解析に基づく改行挿入

従来手法 [3] では，従来手法 [2] と同様の問題設定において，漸進的係り受け解析の結果を ME の素性に利用した改行挿入手法を提案している．従来手法 [3] で使用されている漸進的係り受け解析手法では，文節が入力されるごとに既入力文節の係り先が他の既入力文節のいずれであるか，あるいは，未入力文節であるかを同定し出力する．この出力を元にして，「行頭から文節 b_i までの間で係り受けが閉じているか否か」という素性 f_{12} を判定し，従来手法 [2] の素性に加えて用いている．[1)]

図 1 は，講演の文節列の一部「…/日本が/軍事大国に/ならない/再び/…」中の文節 b_{i+1}「再び」が入力され，文節 b_i「ならない」の直後に改行を挿入するか否かを推定するときに，素性 f_{12} の抽出時に参照する係り受け構造である．文節 b_{i-2}，文節 b_{i-1} が文節 b_i に係ることから，文節 b_{i-2} から文節 b_i までの文節列は係り受けが閉じていると判定する．

1)　なお，係り受け情報 3 種 $f_7 \sim f_9$ についても漸進的係り受け解析の結果を利用して取得している．

表 1　従来手法と提案手法の違い

	素性 F^B	素性 F^A	素性 F^D	RL	モデル
従来手法 [2]	○	×	×	×	ME
従来手法 [3]	○	×	○	×	ME
従来手法 [4]	○	×	×	○	BERT
提案手法	○	○	○	○	BERT

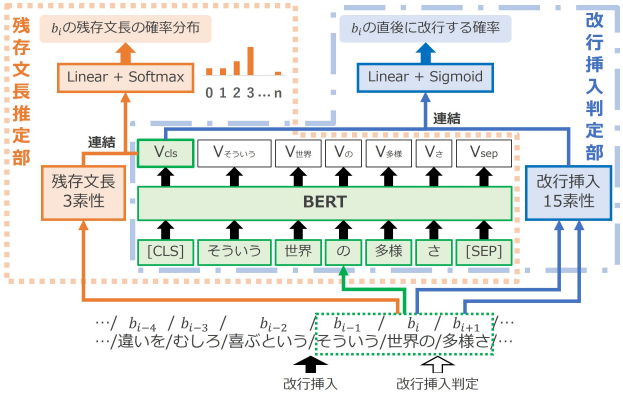

図 2　提案手法の概要

2.3　残存文長を考慮した逐次的改行挿入

従来手法 [4] では，残存文長推定と改行挿入判定を同一の BERT モデルで同時実行する手法を提案している．飯泉ら [4] は河村らの既存研究 [7] と同様に，文 s が n_s 個の文節から成り，文頭から i 番目の文節 b_i まで既に入力されているときの残存文長 $RL(s,i)$ を $RL(s,i) = n_s - i$ により定義している．また，従来手法 [2] と同様の問題設定を用いており，b_i が入力されたときの残存文長 $RL(s,i)$ を同時に推定することで，残存文長を考慮した改行挿入判定を実現している．

3　提案手法

2 章で挙げた各逐次的改行挿入手法では，それぞれ以下の情報を考慮していた．

- **素性 $F^B = \{f_1, \cdots, f_{11}\}$**：2.1 節で述べた従来手法 [2] の基本素性 11 種．
- **素性 $F^D = \{f_{12}\}$**：2.2 節で述べた従来手法 [3] の漸進的係り受け解析結果に基づく素性 1 種．
- **残存文長 RL**：2.3 節で述べた従来手法 [4] が改行挿入判定と同時的に推定する残存文長．

本稿では，これらの情報の他に，一般的な音響情報として声の高さ，大きさ，発話速度に関する各 1 種，計 3 種の素性 $F^A = \{f_1^A, f_2^A, f_3^A\}$ を考慮した手法を提案する．表 1 に，従来手法と提案手法の位置づけを示す．

提案手法では，従来手法 [2] と同様の問題設定を用いており，文節 b_{i+1} が入力されたときに，その直前の文節 b_i との間に改行を挿入するか否かを判定する．残存文長については，従来手法 [4] と同様に，改行挿入判定と同時に残存文長を推定する．

提案手法の概要を図 2 に示す．図 2 は，講演の文節列の一部「…/違いを/むしろ/喜ぶという/そういう/世界の/多様さ/…」中の文節 b_{i+1}「多様さ」が入力されたとき，文節 b_i「世界の」の直後に改行を挿入する確率と残存文長 $RL(s,i)$ の確率分布を推定する様子である．ただし，それまでの改行挿入判定結果において，直近の改行挿入位置は文節 b_{i-2}「喜ぶという」の直後であるとした際の様子である．

提案手法では，それまでの改行挿入判定結果において，直近の行頭文節となった文節（図 2 では「そういう」）から入力された文節（図 2 では「多様さ」）までのサブワード列を BERT への入力として用いる．また，改行挿入部では，従来手法 [3] と同様の 12 種類の素性 F^B, F^D と，音響情報に基づく 3 種類の素性 F^A と合わせた 15 次元のベクトルを，BERT の<CLS>の入力に対応した出力（図 2 では V_{cls}）と連結したものを Linear+Sigmoid に入力することで，改行挿入の確率を出力する．この確率が 0.5 以上の場合には，改行を挿入すると判定する．残存文長推定部では，従来手法 [4] と同様に，b_i の直後にポーズ，フィラー，言い淀みのそれぞれが出現しているか否かを表した 3 次元のベクトルを用意し，これを V_{cls} と連結したものを Linear+Softmax に入力し，その出力を残存文長の確率分布とする．[2)]

3.1 音響情報

改行挿入位置と，声の高さ・大きさ・発話速度との各関係を事前分析し[3)]，音響情報に関する素性を新たに定めた．具体的には，声の高さ，大きさ，発話速度のそれぞれについて，文節 b_{i+1} と文節 b_i の差分を求め[4)]，以下のように，クラス分類したものを素性として用いることとした．

- **素性** f_1^A：声の高さの差が「-600Hz 未満，-600Hz 以上-200Hz 未満，200Hz 以上 600Hz 未満，600Hz 以上」のいずれであるか．
- **素性** f_2^A：声の大きさの差が「-12dB 未満，-12dB 以上 2.4dB 未満，2.4dB 以上 9.6dB 未満，9.6dB 以上」のいずれであるか．
- **素性** f_3^A：発話速度の差が「-3.2 モーラ/秒未満，-3.2 モーラ/秒以上 1.6 モーラ/秒未満，1.6 モーラ/秒以上」のいずれであるか．

表 2 各比較手法のパラメータ

	学習率	バッチサイズ	エポック数
[BERT^{LF}]	$2e-5$	16	11
[BERT^{LF}+F^B]	$2e-5$	16	2
[BERT^{LF}+F^B + F^A]	$9e-6$	16	27
[BERT^{LF}+F^B + F^A + F^D]	$9e-6$	16	21
提案手法	$4e-5$	16	6

4 評価実験

提案手法の有効性を評価するために，日本語講演データを用いて改行挿入実験を行った．

4.1 実験概要

実験データには，同時通訳データベース [8] に収録されている音声データと，日本語講演音声の書き起こしデータを使用した．なお，書き起こしデータの全てのデータに形態素情報，節境界情報，係り受け情報，改行位置が人手で付与されている．実験は，14 講演を用いた交差検定により実施した．すなわち，1 講演をテストデータとし，残りの 13 講演を学習データとする実験を 14 回繰り返した．ただし，14 講演のうち 2 講演については，開発データとして使用するため評価データから取り除き，残りの 12 講演に対して評価を行った．

比較手法として以下を用意した．

- **[BERT^{LF}]**：提案手法において，素性 F^B, F^A, F^D を使用せず，残存文長の同時推定を行わない手法．
- **[BERT^{LF}+F^B]**：提案手法において，素性 F^A, F^D を使用せず，残存文長の同時推定を行わない手法．
- **[BERT^{LF}+F^B + F^A]**：提案手法において，素性 F^D を使用せず，残存文長の同時推定を行わない手法．
- **[BERT^{LF}+F^B + F^A + F^D]**：提案手法において，残存文長の同時推定を行わない手法．

2) 提案手法では確率分布の次元数を開発データにおいて出現した最も長い文節数としている，また学習時は正解の残存文長の one-hot ベクトルを入力している．
3) 事前分析には，4.1 節で述べる開発データを用いた．
4) 各文節の声の高さと大きさは，その音声区間において 8ms ごとに各値を算出し，その平均値とした．

表 3　実験結果

	再現率 (%)	適合率 (%)	F 値
[BERT^{LF}]	74.91	70.64	72.72
[BERT^{LF}+F^B]	75.29	71.72	73.46
[BERT^{LF}+F^B + F^A]	75.52	71.54	73.48
[BERT^{LF}+F^B + F^A + F^D]	76.00	71.30	73.57
提案手法	78.23	71.84	74.90

ここで BERT^{LF} は，改行挿入判定のみを行う BERT モデルを意味する．

各手法では，モデルは PyTorch [5] を用いて実装し，BERT の事前学習モデルは東北大学の公開モデル[6] を用いた．学習アルゴリズムは AdamW を採用した．ハイパーパラメータは，開発データを用いて予備実験し，F 値が最大となった表 2 の値を採用した．素性 F^A の音響情報は librosa [7] を用いて算出した．

評価では，各手法について，同一実験を 5 回繰り返し，再現率，適合率，F 値の平均値を測定した．

4.2　実験結果

実験結果を表 3 に示す．

提案手法は再現率，適合率，F 値のいずれもすべての比較手法を上回った．提案手法と各比較手法の各組合せの F 値を比較したところ，全ての組合せにおいて，有意差が認められた ($p < 0.01$)[8]．提案手法と [BERT^{LF}+F^B + F^A] の間で有意差が認められることから，漸進的係り受け解析結果と残存文長推定とをともに考慮することの有効性を確認した．

一方，比較手法において，F^B，F^A，F^D を順に追加する前後でそれぞれ比較すると，F 値が上昇しているものの，有意差は認められなかった ($p > 0.05$)．各情報の効果の検証は今後の課題である．

4.3　考察

まず，漸進的係り受け解析結果と残存文長推定とをともに考慮することによる好影響を考察する．提案手法が正解し，各比較手法が不正解となった例を図 3 に示す．提案手法では，「国際社会を」と「構成する」の間に正しく改行を挿入できているが，各比較手法では不正解となり「多様な」と「ステイクホルダーとして」の間に余分な改行が挿入されている．漸進的係り受け解析は，「国際社会を」が「構成する」に係るという係り受け構造を出力しており，提案手法は「国際社会を構成する」の係り受けが閉じているという情報を利用できた．残存文長推定の結果は，従来研究 [4, 7] と同様に，確率分布の期待値が「0，1，2〜3，4 以上」のいずれに属するかを求めると，図 3 の例では，残存文長の推定結果は「4 以上」であり，残存文長が長いという情報を利用できた．そのため，入力された文節列の間に構文的なまとまりがあることや文末が近くないことを考慮でき，適切な位置で改行できたものと考えられる．

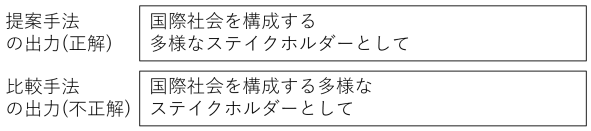

図 3　提案手法が正解，比較手法が不正解の例

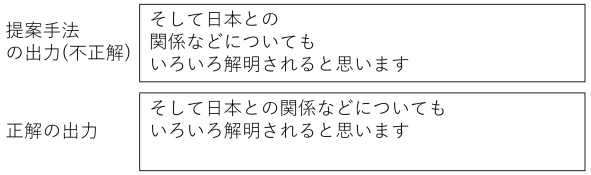

図 4　提案手法が不正解の例

次に，悪影響を考察する．提案手法が不正解となった例を図 4 に示す．提案手法では「日本との」と「関係などについても」の間に，余分な改行が挿入されている．「日本との」と「関係などについても」の間の改行挿入判定時において，漸進的係り受け解析の結果では「そして」が「日本との」に係るとなっており，提案手法は「そして日本との」の係り受けが閉じているという情報を利用することになる．また，残存文長推定の出力確率分布の期待値は「4 以上」であり，残存文長が長いという情報を利用することになる．そのため，当該文節列に構文的なまとまりがあり，文末が近くないということを考慮し，改行挿入すると判定したものと考えられる．漸進的係り受け解析や残存文長推定の性能には限界があり，そのエラーへの対処は今後の課題である．

5　おわりに

本稿では，漸進的係り受け解析と残存文長推定に基づく講演文への逐次的な改行挿入手法を提案した．実験の結果，漸進的係り受け解析結果と残存文長推定をともに考慮することの有効性を確認した．今後は，改行挿入判定の際に，残存文長推定と漸進的係り受け解析を同時に行う深層学習モデルについて検討し，更なる精度向上を図りたい．

5）https://pytorch.org/
6）https://github.com/cl-tohoku/bert-japanese
7）https://librosa.org/
8）評価用データ 12 講演の各講演ごとに F 値を算出し，ウィルコクソンの符号順位検定を行った．なお，本論文の以下で示す統計的検定は全て同じ方法で行った．

謝辞

本研究は，一部，科学研究費補助金基盤研究 (C) No. 19K12127 により実施した．

参考文献

[1] 村田匡輝, 大野誠寛, 松原茂樹. 読みやすい字幕生成のための講演テキストへの改行挿入. 電子情報通信学会論文誌 D, Vol. J92-D, No. 9, pp. 1621–1631, 2009.

[2] 大野誠寛, 村田匡輝, 松原茂樹. 講演のリアルタイム字幕生成のための逐次的な改行挿入. 電気学会論文誌, Vol. 133-C, No. 2, pp. 418–426, 2013.

[3] 大野誠寛, 松原茂樹. 文節間の依存・非依存を同定する漸進的係り受け解析. 電子情報通信学会論文誌 D, Vol. J98-D, No. 4, pp. 709–718, 2015.

[4] 飯泉智朗, 大野誠寛, 松原茂樹. 残存文長を考慮した講演テキストへの逐次的な改行挿入. 言語処理学会 第 28 回年次大会 発表論文集, Vol. 2022, No. 1, pp. 2061–2065, 2022.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **Proceedings of the 2018 Annual Conference of the North American Chapter of the Association for Computational Linguistics**, pp. 4171–4186, 2018.

[6] 大野誠寛, 神谷優貴, 松原茂樹. 対話コーパスを用いた相づち生成タイミングの検出. 電子情報通信学会論文誌 A, Vol. J100-A, No. 1, pp. 53–65, 2017.

[7] 河村天暉, 大野誠寛, 松原茂樹. 漸進的な言語処理のための独話文に対する残存文長の推定. 情報処理学会第 82 回全国大会講演論文集, Vol. 2020, No. 1, pp. 447–448, 2020.

[8] Shigeki Matsubara, Akira Takagi, Nobuo Kawaguchi, and Yasuyoshi Inagaki. Bilingual spoken monologue corpus for simultaneous machine interpretation research. In **Proceedings of the 3rd International Conference on Language Resources and Evaluation**, pp. 153–159, 2002.