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

飯泉 智朗 [1,a)]　大野 誠寛 [1,b)]　松原 茂樹 [2]
[1] 東京電機大学大学院未来科学研究科　[2] 名古屋大学情報連携推進本部
[a)] 20fmi03@ms.dendai.ac.jp　[b)] ohno@mail.dendai.ac.jp

概要

講演を対象とした字幕生成では，講演独特の長い文が複数行にまたがって表示されることになるため，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，残存文長を考慮した逐次的な改行挿入手法を提案する．提案手法の特徴は，BERT を用いて残存文長推定と改行挿入判定を同時に実行する点にある．評価実験の結果，改行挿入の精度向上に対して，提案手法の有効性を確認した．

1　はじめに

聴覚障碍者や高齢者，外国人が講演音声を理解することを支援するため，字幕生成システムが開発されている．講演は，文が長くなる傾向にあり，1 文が複数行にまたがって表示されることになるため，読みやすい字幕を生成するには適切な位置に改行を挿入することが重要となる [1]．

これまでにも字幕テキストに改行を挿入する研究はいくつか行われている [2, 3]．村田ら [2] は，入力文全体を考慮して，最適な改行結果を求める手法を提案している．しかし，1 文全体があらかじめ与えられることを前提としており，講演の進行と同期したリアルタイムでの字幕生成には必ずしも適さない．それに対し大野ら [3] は，入力に対する出力の同時性をより高めた手法として，文節が入力されるたびに，その直前の文節境界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）を用いて逐次的に判定する手法（以下，従来手法 [3]）を提案している．しかしながら，未入力部分の情報を使えないという制約があり，1 文全体で最適化を図る手法と比べて，その精度は低く，改善の余地がある．

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，BERT（Bidirectional Encoder Representations from Transformers）[4] 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する．残存文長とは文の残りの長さを意味する．一般に，文がもう少しで終わる場所での改行の必要性は低下するなど，残存文長と改行位置には関連があると考えられる．本研究では，BERT を用いて，残存文長推定と改行挿入判定とを同時に行うことにより，逐次的な改行挿入における精度向上を試みる．

2　残存文長

残存文長を推定する既存研究として河村ら [5] の研究がある．河村ら [5] は残存文長を，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（すなわち，既入力文節数が i であるとき）の残存文長 $RL(s,i)$ を $RL(s,i) = n_s - i$ により定義している．

本研究においても上記の定義を用いることとする．

2.1　河村ら [5] による残存文長推定

河村ら [5] は残存文長を RNN[6] を用いて推定している．河村らの手法では，1 文（$s = b_1 \ldots b_{n_s}$）を構成する文節が入力されるごとに，文頭から現在入力された文節までの形態素系列（ポーズ，フィラー，言い淀みを表す記号を含む）を RNN に入力し，そのときの残存文長を推定する．この推定を，文節 b_1 が入力されてから文節 b_{n_s} が入力されるまで繰り返す．なお，話し言葉に現れるポーズ，フィラー，言い淀みは，順に記号 P，F，D で汎化して表した後，これらの記号を 1 形態素として扱っている[1)]．河村らの手法では RNN に残存文長の確率分布を出力させ，その期待値（小数点第 1 位を四捨五入）が「0，1，2~3，4 以上」の 4 クラスの中でどのクラスに属するかを求めている．なお，RNN の出力層の次元

1)　ポーズ，フィラー，言い淀みは事前に検出できるとする．

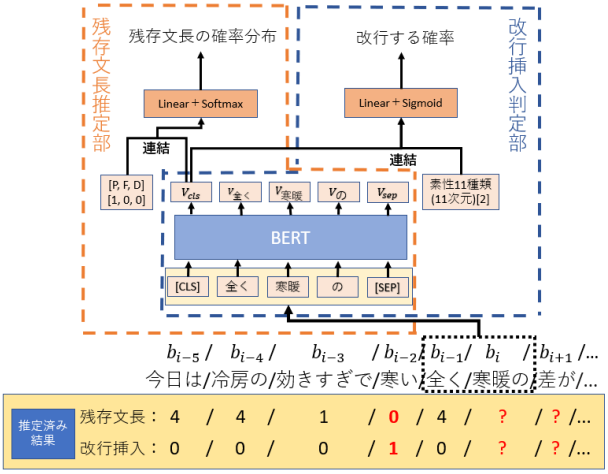

図 1　提案手法の概要

数は，最も長いと想定する文の長さ（文節数）としている．RNN の隠れ層は 1 層の LSTM（Long Short Term Memory）[7] により構成し，RNN への各入力（ポーズ，フィラー，言い淀みの各記号，及び，形態素）は one-hot ベクトルで表現している．

2.2　関連研究

残存文長を推定する研究は，管見の限り河村らの研究のほかに存在しないものの，関連する研究として，テキストの未入力部分を予測する研究がいくつか存在する．例えば，小松ら [8] はテキスト入力の際に，それまでに入力された文脈を考慮して次に入力される確率の高い候補を予測している．また，恒松ら [9] は，音声が入力される場面において，次に発話される単語を予測している．また，Wenyan ら [10] は，リアルタイム翻訳を実現するために，ドイツ語や日本語などにみられる主動詞が最後に来る言語に対して，その最終動詞を予測をしている．

3　残存文長を考慮した改行挿入

本研究では，従来手法 [3] と同じ問題設定とし，1 講演分の文節列が 1 文節ずつ入力されることを想定する．提案手法は，講演の最初から数えて $i+1$ 番目の文節 b_{i+1} が入力されるごとに，b_i が入力された時点の残存文長 $RL(S(i),i)$（$S(i)$ は b_i が属する文）を推定すると同時に，b_i と b_{i+1} の間で改行するか否かを判定する．同一のモデルで改行挿入判定と残存文長推定を実行することにより，残存文長を考慮した改行判定の実現が期待できる．

提案手法の概要を図 1 に示す．図 1 は講演の文節列の一部「... 今日は/冷房の/効きすぎで/寒い/全く/寒暖の/差が/激しいです...」の中の文節 b_{i+1}「差が」が入力されたとき，文節 b_i「寒暖の」の直後に改行を挿入する確率と残存文長 $RL(S(i),i)$ の確率分布を推定する様子である[2]．提案手法では，残存文長推定と改行挿入判定を BERT を用いた同一のモデルで行う．BERT への入力は，b_i より前の残存文長推定結果において，b_i が属する文 $S(i)$ の文頭文節であると推定された文節から，b_i までのサブワード列とする．図 1 では，b_{i-2} の残存文長が 0 文節であると推定された結果[3] に基づいて，b_{i-1} を $S(i)$ の文頭文節とみなす．その結果，BERT への入力は，「全く寒暖の」をサブワード分割した「<CLS> 全く 寒暖 の <SEP>」となる．

残存文長推定部では，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，BERT の<CLS>に対応した出力（図 1 では 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している[4]．改行挿入判定部では，従来手法 [3] で使用された 11 種類の素性を 11 次元のベクトルとして用意し，BERT の<CLS>に対応した出力（V_{cls}）と結合したものを Linear ＋ Sigmoid に入力する．その出力（1 次元）を改行を挿入する確率としている．この確率が 0.5 以上であれば改行を行う．

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 [11] の日本語講演音声書き起こしテキストを使用した．全データに形態素情報，節境界情報，改行位置が人手で付与されている．

実験は，全 16 講演を用いた交差検定によって行った．すなわち，1 講演をテストデータとし，残りの 15 講演を学習データとして改行位置を同定する実験を 16 回繰り返した．ただし，16 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 14 講演に対して

2）　本稿では文節境界をスラッシュで表す．
3）　河村ら [5] と同様に，残存文長の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ている．
4）　本研究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．

表 1　改行挿入判定の実験結果

	再現率 (%)	適合率 (%)	F 値
従来手法 [3]	76.27 (5,489/7,197)	70.00 (5,489/7,841)	73.00
個別手法 (残存文長無)	81.88 (5,893/7,197)	72.06 (5,893/8,178)	76.66
個別手法 (残存文長有)	81.78 (5,886/7,197)	72.45 (5,886/8,124)	76.84
提案手法	83.69 (6,023/7,197)	73.24 (6,023/8,224)	78.11

評価を行った．評価には，正解データの改行位置に対する再現率，適合率，F 値を用いた．

比較手法として従来手法 [3] のほかに，以下の手法を用意した．

個別手法 (残存文長無)：残存文長推定と改行挿入判定を個別に行う手法．具体的には，図 1 の残存文長推定部（赤枠部分）と改行挿入判定部（青枠部分）のモデルをそれぞれ用意し，残存文長推定と，改行挿入判定を独立に行う．ただし，改行挿入モデルには残存文長情報を入力しない．

個別手法 (残存文長有)：残存文長推定と改行挿入判定を個別に行い，改行挿入判定に残存文長情報を使用した手法．個別手法 (残存文長無) との違いは，改行挿入判定モデルにおいて，従来手法 [3] で使用された 11 種類の素性と<CLS>に対応した出力に加えて，残存文長推定モデルで出力された確率分布を連結したものを Linear+Sigmoid に入力している点である．なお，上記 2 つの個別手法における BERT への入力は提案手法と同じである．

モデルは Pytorch を用いて実装した．学習アルゴリズムは SGD を採用した．パラメータの更新はミニバッチ学習（学習率 0.02，バッチサイズ 16）により行い，エポック数は 20 とした．BERT は東北大学が公開している事前学習済み BERT モデル[5)] を用いてファインチューニングを行っている．また，パラメータ更新には改行挿入判定部における BCELoss と残存文長推定部における CELoss の平均を使用した．なお，提案手法で用いる素性のうち，係り受け情報に関しては，従来手法 [3] が ME による推定結果を用いているのに対し，提案手法では，BERT による推定結果を推論時に用いている．

4.2　実験結果

各手法による改行挿入判定の適合率，再現率，F 値を表 1 にそれぞれ示す．提案手法は，すべての評価指標において最高値を達成しており，提案手法の有効性を確認した．

個別手法（残存文長無）と比較して，個別手法（残存文長有）は F 値において上回っているものの，その差はわずかであった．それに対して，提案手法は大幅に上回った．残存文長を考慮する際に，個別に推定した残存文長情報を単に利用するのではなく，同時に推定することが有効であることを確認できる．

5) https://github.com/cl-tohoku/bert-japanese

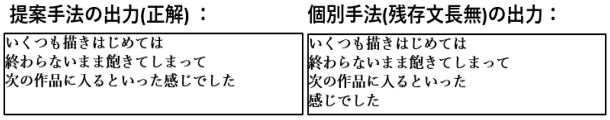

図 2　提案手法が正解，個別手法 (残存文長無) が不正解の例

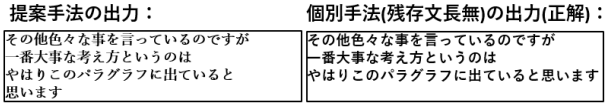

図 3　個別手法 (残存文長無) が正解，提案手法が不正解の例

表 2　残存文長が正解した場合と不正解の場合の実験結果

	再現率 (%)	適合率 (%)	F 値
正解	85.37(4,143/4,853)	76.76(4,143/5,397)	80.84
不正解	80.20(1,880/2,344)	66.50(1,880/2,827)	72.71
合計	83.69(6,023/7,197)	73.24(6,023/8,224)	78.11

5　考察

5.1　残存文長を考慮しない手法との比較

本節では，残存文長の推定結果を考慮することによる影響を考察する．提案手法が正解し，個別手法 (残存文長無) が不正解となった例を図 2 に示す．提案手法の改行位置は正解データと完全に一致しているが，個別手法 (残存文長無) では「入るといった」の後で余分に改行されている．「入るといった」の文節が入力された際に推定された残存文長の確率分布の期待値は約 1 であり，実際の残存文長と一致していた．その結果，当該情報を利用している提案手法では，文がもう少しで終わることを考慮でき，余分な改行を抑えることに成功したものと考えられる．

次に，個別手法 (残存文長無) が正解し，提案手法が不正解となった例を図 3 に示す．正解の改行位置と比べて，提案手法の出力では，「出ていると」の直後に余分な改行が行われている．「出ていると」の

表 3　残存文長推定の実験結果（P：適合率，R：再現率，F：F 値）

	残存文長：0 文節			残存文長：1 文節			残存文長：2〜3 文節			残存文長：4 文節以上		
	R(%)	P(%)	F	R(%)	P(%)	F	R(%)	P(%)	F	R(%)	P(%)	F
河村らの手法 [5]	88.45 (1,516/1,714)	87.43 (1,516/1,734)	87.94	31.76 (540/1,700)	27.33 (540/1,976)	29.38	29.13 (960/3,296)	22.93 (960/4,186)	25.66	72.08 (10,089/13,997)	78.75 (10,089/12,811)	75.27
個別手法	86.58 (1,484/1,714)	91.21 (1,484/1,627)	**88.84**	37.76 (642/1,700)	23.05 (642/2,785)	28.63	28.55 (941/3,296)	20.99 (941/4,484)	24.19	66.57 (9,318/13,997)	78.89 (9,318/11,811)	72.21
提案手法	85.01 (1,457/1,714)	92.10 (1,457/1,582)	88.41	32.00 (544/1,700)	29.74 (544/1,829)	**30.83**	28.88 (952/3,296)	23.13 (952/4,116)	**25.69**	74.43 (10,418/13,997)	79.04 (10,418/13,180)	**76.67**

文節が入力された際に推定された残存文長の確率分布の期待値は約 5 であったが，実際の残存文長は 1 であり，大きく外れていた．その結果，提案手法は，まだ文が長く続くという情報を考慮することになり，この位置で過剰に改行することになったと考えられる．

表 2 は，残存文長の推定結果（確率分布の期待値を 4 クラス分類したもの）が正解していたか否かによって場合分けし，各場合における提案手法の再現率，適合率，F 値を再評価した結果である．残存文長を正しく推定できている場合は，そうでない場合と比べて，いずれの評価指標においても大幅に上回っていることがわかる．

以上は，残存文長推定の精度が高まれば改行挿入判定の精度が向上することを示唆しており，改行挿入において残存文長を考慮することの有効性を確認できる．

5.2　残存文長推定の実験結果

本節では，4 節の実験結果に基づいて，提案手法による残存文長推定の精度を評価する．評価では，河村らの研究 [5] と同様に，残存文長の確率分布の期待値（小数点第 1 位を四捨五入）を「0，1，2~3，4 以上」の 4 クラス分類し，各クラスの再現率，適合率，F 値を測定する．

比較手法として，4 節の個別手法に加えて，河村らの手法 [5] を用意した．ただし，河村らの手法 [5] では 1 文ごとに推定を行い，文末を既知として扱っているのに対し，提案手法では，1 講演ごとに推定を行い，残存文長が 0 文節と推定された文節を文末として扱っているため，問題設定は異なっている．

各手法による残存文長推定の適合率，再現率，F 値を表 3 にそれぞれ示す．河村らの手法 [5] と，提案手法を比較するとすべてのクラスの F 値において上回っていることがわかる．次に個別手法と提案手法を比較すると，0 文節以外のすべてのクラスにおいて上回っており，特に 4 文節以上のクラスに至っては約 4 ポイントの上昇を記録している．提案手法では，残存文長推定との同時学習によって，改行挿入判定の精度向上を目指していたが，残存文長推定においても精度が上昇することがわかった．

表 4　人間による推定との比較（0：0 文節，1：1 文節，2〜3：2〜3 文節，4〜：4 文節以上）

	F 値 (%)			
	0	1	2〜3	4〜
人間（平均）	84.51	30.60	26.90	64.68
人間（最低）	79.82	31.48	27.37	48.57
人間（最高）	**88.46**	38.84	27.29	77.87
提案手法	88.08	**39.38**	**27.92**	**79.69**

5.3　人間による残存文長推定との比較

本節では，提案手法と人間による残存文長推定 [5] とを比較する．河村らの研究 [5] では，4 節の実験データの中から 100 文を抽出し，各文の各文節が入力されるごとに 8 人の人間が残存文長を推定する被験者実験を行っている．その 100 文に対して提案手法を適用した結果を用いて比較評価する．

表 4 に，提案手法と人間による推定の F 値を示す．すべてのクラスにおいて提案手法が人間（平均）を上回った．人間（最高）の F 値との比較では，0 文節のクラスでわずかに下回ったのみで他のクラスでは上回った．以上は，提案手法が，人間と同程度，若しくは同程度以上に，残存文長を推定できる可能性を示唆している．

6　おわりに

本稿では，講演テキストを対象に，残存文長を考慮した逐次的な改行挿入手法を提案した．実験の結果，提案手法は，従来手法や個別手法よりも再現率，適合率，F 値において上回っており，その有効性を確認した．また，逐次的な改行挿入において残存文長を考慮することの有効性を確認した．

今後は，残存文長情報のより効果的な導入方法について検討し，さらなる精度向上を図りたい．

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] 中野聡子, 金澤貴之, 牧原功, 黒木速人, 上田一貴, 井野秀一, 伊福部達. 聴覚障害者向け音声同時字幕システムの読みやすさに関する研究 : (1)-改行効果に焦点をあてて. ヒューマンインタフェース学会論文誌, Vol. 10, No. 4, pp. 435–444, 2008.

[2] 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, Vol. J92-D, No. 9, pp. 1621–1631, 2009.

[3] 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.

[4] 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.

[5] 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.

[6] Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. Recurrent neural network based language model. In **Proceedings of the 11th Annual Conference of the International Speech Communication Association**, Vol. 2, pp. 1045–1048, 2010.

[7] Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. LSTM neural networks for language modeling. In **Proceedings of the 13th Annual Conference of the International Speech Communication Association**, pp. 194–197, 2012.

[8] 小松弘幸, 高林哲, 増井俊之. 動的略語展開を利用した文脈をとらえた予測入力. 情報処理学会論文誌, Vol. 44, No. 11, pp. 2538–2546, 2003.

[9] 恒松和輝, 中村哲. 入力音声に続く文章の予測. 情報処理学会研究報告, Vol. 2019-NL-241, No. 27, pp. 1–4, 2019.

[10] Wenyan Li, Alvin Grissom II, and Jordan Boyd-Graber. An attentive recurrent model for incremental prediction of sentence-final verbs. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings**, pp. 126–136, 2020.

[11] 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.