E-014

読みやすい字幕生成のための講演テキストに対する改行挿入の改善
Improvement of Linefeed Insertion into Lecture Transcription for Automatic Captioning

飯泉 智朗[†,a)]　大野 誠寛[†,b)]　松原 茂樹[‡]
Tomoaki Iizumi　Tomohiro Ohno　Shigeki Matsubara

1　はじめに

聴覚障害者や高齢者，外国人らによる講演音声の理解を支援するための技術として，字幕生成システムの開発が望まれている．字幕を生成する際，提示されたテキストが読みやすくなるように，適切な位置に改行が挿入されている必要がある．

これまでに著者らは，読みやすい字幕を生成するための要素技術として，RNN[1] を用いた日本語講演テキストへの改行挿入手法（以下，従来手法）を提案している [2]．本稿では，従来手法の実験結果を分析し，不適切な改行を抑制するため，新たな素性の追加を試みる．日本語講演テキストを対象とした改行挿入実験の結果，本稿で追加した素性の有効性を確認した．

2　従来研究 [2]

従来手法 [2] では，形態素解析，文節まとめ上げ，節境界解析，係り受け解析が施された文を入力とし，入力文中の各文節境界に対して，その位置に改行をするか否かを同定する．なお，1 行の最大文字数を 20 文字と設定し，各行の文字数がそれ以下となるようにしている．

従来手法では，1 文中に挿入され得る改行位置のすべての組合せの中から，最適な組合せを確率モデルを用いて決定する．すなわち，入力文の文節列を $B = b_1 ... b_n$ とするとき，$P(R \mid B)$ を最大にする改行挿入結果 $R = r_1 ... r_n$ を動的計画法により求める．r_i は文節 b_i の直後に改行が挿入されるか $(r_i = 1)$ 否か $(r_i = 0)$ のいずれかの値をとる．

$P(R \mid B)$ は，各文節境界に改行が挿入されるか否かは直前の改行位置を除く，他の改行位置とは独立であると仮定し，$P(R \mid B) = \prod_{i=1}^{n} P(r_i \mid R_{k_i}^{i-1}, B)$ により求める．$P(r_i \mid R_{k_i}^{i-1}, B)$ は，入力の文節列 B が与えられ，文節 b_i の直前の改行位置が同定されているときに，b_i の直後に改行が挿入される，または，挿入されない確率を表す．文節 b_i の直前の改行位置（b_i が表示される 1 つ前の行の行末位置）の文節を $b_{k_i} (k_i < i)$ とし，$R_{k_i}^{i-1}$ は，b_{k_i} から b_{i-1} までの改行結果 $R_{k_i}^{i-1} = r_{k_i} r_{k_i+1} \cdots r_{i-1} = 10 \cdots 0$ を表す．

$P\left(r_i \middle| R_{k_i}^{i-1}, B\right)$ は RNN により推定する．具体的には，文節列 $B_{k_i+1}^{i} = b_{k_i+1} b_{k_i+2} \cdots b_i$ から得られる素性の系列 $F = F_{k_i+1} F_{k_i+2} \cdots F_i$ を RNN に入力し，$P\left(r_i \middle| R_{k_i}^{i-1}, B\right)$ の 2 値 $(r_i = 0 \text{ or } 1)$ の確率分布を得る．ここで，F_i は文節 b_i から得られる素性列を意味し，

$$F_i = \left(f_{i,j}^1 f_{i,j}^2\right)^{1 \le j \le m_i} f_i^3 f_i^4 f_i^5 f_i^6 f_i^7 f_i^8$$

として表される．素性 f^1 から f^8 は順に，語の出現形，語の品詞細分類，文頭からの文節番号，係り先文節までの距離，直前の文節から係られているか，連体節から係られているか，節境界の種類，ポーズの有無である．f^1 と f^2 は，文節 b_i が m_i 個の形態素により構成されるとすると，各形態素から順番に抽出されるため，

$$\left(f_{i,j}^1 f_{i,j}^2\right)^{1 \le j \le m_i} = f_{i,1}^1 f_{i,1}^2 f_{i,2}^1 f_{i,2}^2 \cdots f_{i,m_i}^1 f_{i,m_i}^2$$

が RNN に入力される．なお，素性 $f_{i,j}^1$ は，文節 b_i の j 番目の形態素から得られる素性 f^1 を意味する．

従来研究 [2] では，本稿 5 節と同じ実験が実施され，従来手法の性能（表 1 の最終行[*1]）が報告されている．

表 1　従来手法の実験結果 (節境界の有無に基づく再評価含む)

	再現率 (%)	適合率 (%)	F 値
節境界有	93.64 (3,237/3,457)	83.58 (3,237/3,873)	88.32
節境界無	74.75 (1,525/2,040)	69.79 (1,525/2,185)	72.19
全体	86.63 (4,762/5,497)	78.61 (4,762/6,058)	82.42

正解：様々なものが外から日本へ
流入してくるということなんですね

従来手法の出力：様々なものが
外から
日本へ流入してくるということなんですね

図 1　従来手法の出力例（小刻みな改行）

正解：あそここう行けというふうに
こういちいち言わないと
行ってくれない訳ですね

従来手法の出力：あそここう行けというふうに
こう
いちいち言わないと行ってくれない訳ですね

図 2　従来手法の出力例（節境界無の場合の誤改行）

3　従来手法のエラー分析

表 1 の最終行を見ると，従来手法による全体改行数は 6,058 であり，正解の全体改行数 5,497 よりも 500 以上多く改行されていることがわかる．また，正解データの行ごとの平均文字数は 13.2 文であるのに対し，従来手法は 12.3 文字であり，短い行が頻出していた．この典型例を図 1 に示す．この例では「様々なものが」と「外から」の直後で小刻みに改行されていることがわかる．このように，まだ文字を埋めるスペースが当該の行に十分残されているのに改行される例が多く見られた．従来手法では，語や文節の文字数に関する素性を使っていないため，このような結果が生じたものと考えられる．

次に，節境界に着目して，従来手法の実験結果を分析した．表 1 に，各文節の直後に節境界があるか否かにより区別して再評価した結果を示す．節境界があると改行されやすい傾向にあり，従来手法でも高い精度を実現できている．しかし，節境界がない場合は，その傾向をつかむことは難しく，その精度は大きく下回っていることがわかる．その典型例を図 2 に示す．従来手法は，「こう」

† 東京電機大学大学院未来科学研究科, Graduate School of Science and Technology for Future Life, Tokyo Denki University.
‡ 名古屋大学情報連携推進本部, Information and Communications, Nagoya University.
a) 20fmi03@ms.dendai.ac.jp
b) ohno@mail.dendai.ac.jp

*1　表 1 は，本稿で再実験した結果であり，文献 [2] の実験結果と比べて，F 値が若干高い値となっている．

で改行しているが，その直後の「言わないと」の直後に節境界「引用節」があり，正解では「言わないと」の後で改行している．このような例を正解するためには，改行判定位置の後方にある節境界の情報を考慮する必要があると考えられる．

4 提案手法

本節では，前節の分析結果に基づき，従来手法に新たな素性を追加した手法を提案する．追加する素性は，前節で指摘した観点ごとに整理し，以下のように決定した．

まず，小刻みな改行を抑制するため，行頭からの文字数情報を利用することとし，以下の素性 f^9 を追加した．

f^9:行頭からの文字数が 2 文字以下であるか

なお，3 文字以上 7 文字以下，7 文字以上など異なる文字数も試したが，予備実験で最も高い効果を示した上記 f^9 を採用した．

次に，直後に節境界のない文節に対する改行判定の精度向上を目指し，以下の素性 f^{10} と f^{11} を追加した．

f^{10} :係り先文節の節境界の種類*2
f^{11} :後方にある直近の節境界との距離

これらの追加は，節境界がない文節に対する改行判定において，その後方に位置する節境界の情報が重要であるという前節の分析に基づくものである．

これら f^9 から f^{11} は文節ごとに抽出される素性であるため，2 節で述べた素性列 F_i は，本稿の提案手法では以下となる．

$$F_i = \left(f_{i,j}^1 f_{i,j}^2\right)^{1\le j\le m_i} f_i^3 f_i^4 f_i^5 f_i^6 f_i^7 f_i^8 f_i^9 f_i^{10} f_i^{11}$$

5 改行挿入実験

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従来研究 [2] と同様の改行挿入実験を行った．

5.1 実験概要

実験データには，同時通訳データベース [4] の日本語講演音声書き起こしテキストを使用した．なお，全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 [3].

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．また，比較は従来手法 [2] と行った．

RNN 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オンライン学習 (学習率 0.01) により行い，更新時にユニットを 0.1 の確率でドロップアウトさせた．エポック数は 4 とした．入力層の入力ベクトル，すなわち one-hot ベクトルのサイズの平均は 5574.1 であった．入力層の出力ベクトルの次元数を 1300，隠れ層（LSTM[5]，1 層）の出力ベクトルの次元数を 400 とした．この値は隠れ層と入力層を 100 から 1500 まで 100 刻みで変化させ開発データを用いた実験において最も F 値が高かったものを採用した．なお，出力層の次元数は 2 とした．

*2 節境界がない場合は無と入力する

表 2 実験結果

	再現率 (%)	適合率 (%)	F 値
従来手法	86.63 (4,762/5,497)	78.61 (4,762/6,058)	82.42
提案手法	86.96 (4,780/5,497)	80.38 (4,780/5,947)	83.54

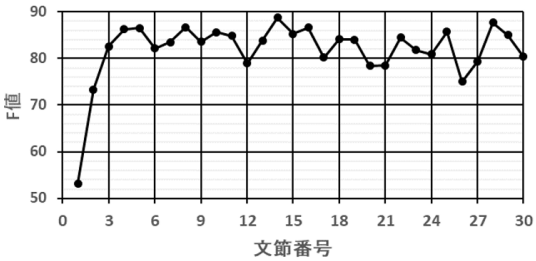

図 3 文節番号ごとの F 値

5.2 実験結果

提案手法及び従来手法 [2] の適合率，再現率，F 値を表 2 にそれぞれ示す．提案手法は，従来手法と比較して，再現率，適合率，F 値がそれぞれ，0.33%，1.77%，1.12 上回っており，提案手法の有効性を確認した．特に，提案手法の総改行数は，従来手法と比べ，100 以上減っており，無駄な改行が抑えられていることがわかる．

図 2 に文節番号ごとの F 値を示す．文節番号とは文頭から，その文節が何文節目かを表した番号である．1 文節目，すなわち，文頭文節に対する F 値は 53.78 であり，他と比べ大きく低下していることがわかる．文頭文節は，前の文との接続の役割を果たすものである場合が多く，その場合，ある程度の強い意味的な切れ目となるといえるため，文頭文節の直後で改行される傾向が強いと考えることができる．しかし，文頭文節の後に，並列関係など，より強い意味的なまとまりが続くと，文字数との兼合いも考慮され，改行されない場合も多いことが確認された．文頭文節の改行判定には，意味的なまとまりの強弱や 1 行当たりの文字数などを総合的に判断することが必要といえ，現状の素性では，これらを捉えることができていないと考えられる．

6 おわりに

本稿では，講演テキストへの改行挿入の精度向上を目指して，従来手法による実験結果を分析し，その改善手法を提案した．実験の結果，提案手法は F 値において，従来手法 [2] の 82.42%を上回る 83.54%を達成しており，本稿で追加した素性の有効性を確認した．今後は，文頭文節に対する改行挿入の精度向上を図る予定である．

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]T. Mikolov et al.,“Recurrent Neural Network Based Language Model,” Proc. INTERSPEECH 2010, pp. 1045–1048, 2010.
[2]飯泉ら,“読みやすい字幕生成のための RNN を用いた講演テキストへの改行挿入,”情報処理学会第 82 回年次大会発表論文集, vol. 2, pp. 451–452, 2020.
[3]村田ら,“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,”信学論, J92-D(9), pp. 1621–1631, 2009.
[4]S. Matsubara et al.,“Bilingual Spoken Monologue Corpus for Simultaneous Machine Interpretation Research,” Proc. LREC 2002, pp. 153–159, 2002.
[5]M. Sundermeyer et al.,“LSTM Neural Networks for Language Modeling,” Proc. INTERSPEECH 2012, pp. 194–197, 2012.