2S-02

残存文長を考慮した講演テキストへの改行挿入

岩村　由香[†a)]　　大野　誠寛[†b)]　　松原　茂樹[‡]

東京電機大学未来科学部[†]　　名古屋大学情報連携推進本部[‡]

1　はじめに

聴覚障碍者，高齢者，外国人らが講演を聴講する際の理解を支援するために自動字幕生成システムの開発が望まれている．しかし講演文は一文が長くなる傾向があることから一文が複数の行にまたがって表示され読みにくくなることが多い．これを読みやすく表示するために適切な位置に改行を挿入する必要がある．

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．残存文長とは文の残りの長さを意味する[2]．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．そのため本手法では，改行挿入の逐次的な判断において，残存文長の情報を用いることにより，改行挿入精度の向上を試みる．

2　従来の逐次的な改行挿入手法

従来手法[1]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機械学習を用いて逐次的に行う．その機械学習には，b_iの主辞，語形などの形態素情報 4 種，節境界情報 2 種，係り受け情報 3 種，行頭からの文字数 1 種，ポーズ情報 1 種類の，合計 11 種類の素性を使用している．ただし，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 文字と設定し，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，機械学習の判定結果に関わらず強制的に改行を挿入する．

3　残存文長推定を用いた改行挿入

本研究では従来手法[1]と同じく形態素情報，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を入力とし，文節が入力されるごとに改行を挿入するか否かの判定を機械学習を用いて逐次的に行う．ただし，従来手法[1]では機械学習に最大エントロピー法を用いているが本研究では SVM を用いた．また，従来手法で用いられている 11 種類の素性に加えて新たに残存文長に関する素性を追加する．

残存文長とは，河村ら[2]の定義と同じく，文節が入力されるごとの文の残りの長さを指す．例えばn個の文節から成る文において，文頭からx番目の文節b_xまで既に入力されている時，残存文長は$n-x$である．河村ら[2]は，文節が入力されるごとに，残存文長の確率分布を RNN により推定し，その確率分布の期待値をその時点での残存文長としている．また，その期待値を 4 クラス(0 文節，1 文節，2〜3 文節，4 文節以上)に区分し，その分類精度を評価している．提案手法では，文節が入力されるたびに河村らの手法[2]が推定した残存文長を用いて，「残存文長が上述の 4 クラスのいずれであるか」という素性を追加する．

4　評価実験

提案手法の有効性を評価するために，日本語講演データに対し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 [1] に収録されている日本語講演音声の書き起こしデータを使用した．なお，全データに形態素情報，係り受け情報，節境界情報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り行った．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挿入の推定を行った．ただし，16 講演のうち 2 講演を開発データとして使用したため評価データから取り除き，残りの 14 講演に対し評価を行った．また，学習時の残存文長の素性値はコーパスから得られる正しい値を用い，テスト時の当該素性値は河村らの手法[2]が推定した値を用いた．

評価では適合率，再現率，F 値を測定した．比較のために，機械学習を SVM に変更して再現した従来手法(提案手法から残存文長の素性を除いた手法)を用意した．SVM には LIBSVM[2] を用い，オプションとして，従来手法には“-c 7.1 -g 0.05”を，提案手法には“-c 33 -g 0.01”に設定した．

[1] http://sidb.jp/

[2] http://www.csie.ntu.edu.tw/~cjlin/libsvm

Linefeed Insertion into Lecture Transcription Considering Remaining Sentence Length

Yuka Iwamura[†,a)], Tomohiro Ohno[†,b)], Shigeki Matsubara[‡]

† School of Science and Technology for Future Life, Tokyo Denki University.

‡ Information and Communications, Nagoya University.

a) 17fi018@ms.dendai.ac.jp

b) ohno@mail.dendai.ac.jp

表 1　実験結果

	従来手法	提案手法
適合率	70.65(5,372/7,604)	71.28(5,510/7,703)
再現率	74.64(5,372/7,197)	76.56(5,510/7,197)
F 値	72.59	73.83

4.2　実験結果

提案手法及び従来手法の適合率，再現率，F 値を表 1 にそれぞれ示す．提案手法は従来手法と比較してF値において1.22%上回っており，残存文長を考慮した提案手法の有効性を確認した．

5　考察

提案手法は，従来手法と同様に，文節b_iの直後で改行しなければ最長文字数を超える場合，それまでに決定した改行有無の判断を覆すことはなく，b_iの直後で強制的に改行する（以下，強制改行）．提案手法のエラー分析をしたところ，このような強制改行に間違いが多いことが明らかとなった．具体例を図 1 の上部に示す．「米国の」や「協議しているということが」の直後に正解とは異なる改行が挿入されている．

そこで以下では，最長文字数制約に伴う改行処理において，強制改行するのではなく，それまでに決定した改行しないとの判断を一部覆すことを許し，より適切な改行位置を遡って探索し改行（以下，バックトラック改行）するように変更し，精度改善を試みる．具体的には，ある行の行頭文節b_iから，b_{i+j}までの直後には改行を挿入しないという判定が行われ，その 1 行（b_iからb_{i+j+1}まで）の文字列が最長文字数を超える場合，b_iからb_{i+j}までの各文節の直後に改行を挿入する確率をそれぞれ算出・記憶しておき，その確率が最大となる文節の直後に改行を挿入するように変更する．なお，上記の処理は最長文字数制約を満たすまで繰り返すものとする．

図 1 の下部に，バックトラック改行結果の例を示す．2 行目の行頭文節「インド側の」から「米国の」まで改行を挿入しないと一旦判定され，「インド側の」から「議会はですね」が 1 行に表示されることになり，2 行目が最長文字数制約を満たさない状況が生じたため，バックトラック改行が行われている．その結果，「インド側の」から「米国の」までの各文節の直後に改行を挿入する確率が最も高い文節である「申し出に対して」の直後に改行が挿入されている．

最長文字数制約に伴うバックトラック改行による精度改善への有効性を確認するため，4 節と同じ設定で，再実験を行った．4 節の実験との違いは，改行挿入手法のみである．具体的には，残存文長の有効性についても再度検証するため，

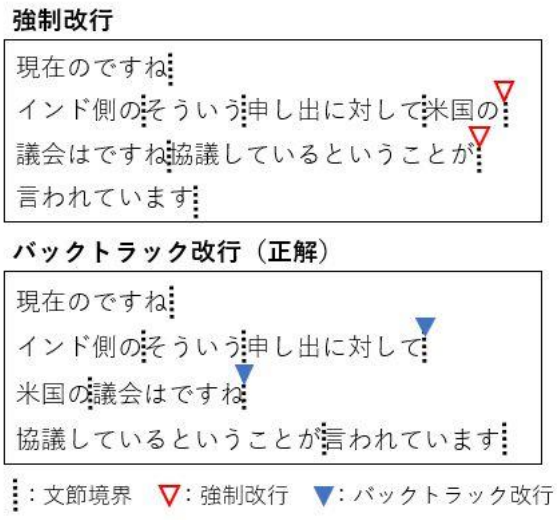

図 1　強制改行とバックトラック改行の例

表 2　バックトラック改行に変更後の実験結果

	従来手法(バックトラック改行)	提案手法(バックトラック改行)
適合率	73.90(5,914/8,003)	74.03(5,989/8,070)
再現率	82.17(5,914/7,197)	83.22(5,989/7,197)
F 値	77.82	78.35

表 1 の従来手法と提案手法において，最長文字数制約に伴う処理をバックトラック改行に変更した手法をそれぞれ用意し改行挿入実験を行った．

再実験の結果を表 2 に示す．表 1 と比較すると，バックトラック改行に変更した手法はいずれも，F 値を上回っており，バックトラック改行の有効性が確認できる．また，最長文字数制約に伴う処理をバックトラック改行に変更した場合においても，残存文長を考慮した提案手法は従来手法と比較してF値が上回っており，残存文長は改行挿入の推定に有効であることが確認できる．

6　まとめ

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，提案手法の有効性を確認できた．今後は，推定された残存文長の確率分布をより効果的に用いて，改行挿入の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] 大野ら，“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 C，Vol.133，No.2，pp.418-426，2013.
[2] 河村ら，“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，No.2，pp.447-448，2020.