4T-03

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

飯泉智朗[†,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統括本部[‡]

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．講演では一文が長くなる傾向にあり，多くの文がスクリーン上で複数行にまたがって表示されることになるため，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．

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，RNN[1]を用いた日本語講演テキストへの改行挿入手法を提案する．日本語講演テキストを用いた改行挿入実験の結果，本手法は従来手法[2]と比べてより適切な位置に改行を挿入できることを確認した．

2　従来の改行挿入手法

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している．なお，字幕を表示するディスプレイの大きさを想定した 1 行における最長文字数を 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．また，日本語における文節は，意味のまとまりの基本単位であることを考慮し，文節境界を改行位置の候補としている．

従来手法では，入力文に対する適切な改行位置を同定するために，1 文中に挿入され得る改行位置のすべての組み合わせの中から，最適な組み合わ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$B = b_1 \dots b_n$とするとき，$P(R|B)$を最大にする改行挿入結果$R = r_1 \dots r_n$を動的計画法により求める．なお，r_iは文節b_iの直後に改行が挿入されるか$(r_i = 1)$か否か$(r_i = 0)$のいずれかの値をとる．

また従来手法では，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，ほかの改行位置とは独立であると仮定することにより，確率$P(R|B)$を$P\left(r_i \middle| R_k^{i-1}, B\right)$の積$P(R|B) = \prod_{i=1}^{n} P\left(r_i \middle| R_k^{i-1}, B\right)$により求めている．$P\left(r_i \middle| R_k^{i-1}, B\right)$は，1 文の文節列$B$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．ここで、文節$b_k (k < i)$は，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）の文節とし，R_k^{i-1}は，b_kからb_{i-1}までの改行結果$R_k^{i-1} = r_k r_{k+1} \cdots r_{i-1} = 10 \cdots 0$を表す．

従来手法では，$P\left(r_i \middle| R_k^{i-1}, B\right)$を最大エントロピー法により算出している．その際には，全部で 15 種類の素性（形態素情報，係り受け情報，行長情報，ポーズ情報，文節の第一形態素の情報）を使用している．

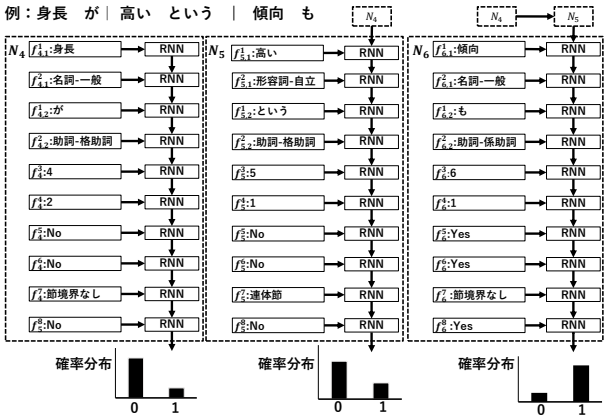

図 1　RNN による$P\left(r_i \middle| R_k^{i-1}, B\right)$の算出例

表 1　素性一覧

ID	素性名	素性値
f^1	語の出現形*1	私，です，世界，ある，など
f^2	語の品詞細分類	名詞-一般，動詞-自立，など 46 種
f^3	文頭からの文節番号	1, 2, 3, など
f^4	係り先文節との距離	0, 1, 2, 3, 4, 5 以上の 6 種
f^5	直前の文節から係られているか	2 値（Yes , No）
f^6	連体節から係られているか	2 値（Yes , No）
f^7	節境界の種類	引用節，連体節，など 90 種
f^8	ポーズの有無	2 値（Yes , No）

3　RNN を用いた改行挿入

本研究では，従来手法[2]の問題設定及び改行挿入アルゴリズムをそのまま採用するが，$P\left(r_i \middle| R_k^{i-1}, B\right)$の算出において，最大エントロピー法の代わりに，RNN を用いることで性能向上を図る．

本手法では，$P\left(r_i \middle| R_k^{i-1}, B\right)$を RNN により推定する際，文節列$B_{k+1}^{i} = b_{k+1} b_{k+2} \cdots b_i$から得られる素性の系列$F = F_{k+1} F_{k+1} \cdots F_i$を RNN に入力し，$P\left(r_i \middle| R_k^{i-1}, B\right)$の 2 値($r_i$=0 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の詳細を表 1 に示す．これらは従来研究[2]を参考に設定したものである．

素性f^1とf^2は，形態素から得られる素性であり，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，

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

が RNN に入力される．なお，素性$f_{i,j}^1$は，文節b_iのj番

*1　語の品詞分類が「固有名詞」であった場合，語の出現形を入力せず「品詞細分類+出現形の文字数」を一つの異なり語として入力している．

Linefeed Insertion into Lecture Transcription Using RNN for Automatic Captioning
Tomoaki Iiizumi[†,a)] ,Tomohiro Ohno[†,b)] , Shigeki Matsubara[‡]
† School of Science and Technology for Future Life, Tokyo Denki University
‡ Information and Communications, Nagoya University
a) 16fi006@ms.dendai.ac.jp
b) ohno@mail.dendai.ac.jp

目の形態素から得られる素性f^1を意味する．一方，素性f^3からf^8は文節から得られる素性であり，例えば，素性f_i^3は，文節b_iから得られる素性f^3を意味する．

例として，7個の文節から成る文「また|そういった|人たちのほうが|身長が|高いという|傾向も|ある」への改行挿入を考える．今，1 文中に挿入され得る改行位置のすべての組み合わせのうち，$R = 0010011$の場合の$P(R|B) = \prod_{i=1}^{n} P\left(r_i \middle| R_k^{i-1}, B\right)$を計算することになるため，$\prod_{i=1}^{n} P\left(r_i \middle| R_k^{i-1}, B\right)(i = 4,5,6)$を RNN により求めるとする。このときの RNN の入出力の概要を図 1 に示す．なお，図 1 のN_iは$P\left(r_i \middle| R_k^{i-1}, B\right)$を求める際の RNN とする．

まず，$P(r_4 = 0|R_3^3, B)$を求める．このときの入力系列$F = F_4$は，b_4が 2 個の形態素から成るため，$F_4 = f_{4,1}^1 f_{4,1}^2 f_{4,2}^1 f_{4,2}^2 f_4^3 f_4^4 \cdots f_4^8$となり，RNN が出力する$P(r_4|R_3^3, B)$の確率分布から，$P(r_4 = 0|R_3^3, B)$を得る．

次に，$P(r_5 = 0|R_3^4, B)$を求める．このときの入力系列は$F = F_4F_5$となり，上記F_4を入力した後，$F_5 = f_{5,1}^1 f_{5,1}^2 f_5^3 f_5^4 \cdots f_5^8$を入力する．この RNN が出力する確率分布から，$P(r_5 = 0|R_3^4, B)$を得る．

最後に，$P(r_6 = 1|R_3^5, B)$を求める．このときの入力系列は$F = F_4F_5F_6$となり，上記F_4とF_5を順に入力した後，$\mathrm{F}_6 = \mathrm{f}_{6,1}^1 \mathrm{f}_{6,1}^2 \mathrm{f}_{6,2}^1 \mathrm{f}_{6,2}^2 \mathrm{f}_6^3 \mathrm{f}_6^4 \cdots \mathrm{f}_6^8$を入力する．この RNN が出力する確率分布から，$P(r_6 = 1|R_3^5, B)$を得る．

4 改行挿入実験

本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[3]の中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[2].

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と適合率を用いる．再現率は正解の改行位置に改行が挿入された割合，適合率は挿入した改行のうち正解と一致する割合である．

本手法との性能比較のために，従来手法[2]による同じ設定での実験結果を用意した．

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 5461.8 であった*2．入力層の出力ベクトルの次元数を 1300，隠れ層（LSTM[4]，1 層）の出力ベクトルの次元数を 100 とした．この値は隠れ層と入力層を 100~1300 まで 100 刻みで変化させ開発データの推定において最も F 値が高かったものを採用した．さらに，出力は 2 値（改行する or 改行しない）の確率分布とするため，出力層の次元数は 2 とした．

*2 交差検定を行ったため，評価に用いた 14 講演に対する全 14 回の実験における平均を求めた．

表 2 実験結果

	再現率	適合率	F 値
従来手法	82.66% (4,544/5,497)	80.24% (4,544/5,633)	81.43
本手法	87.54% (4,823/5,497)	77.69% (4,812/6,194)	82.39

正解：

ですからインポートって
本日のお手元の資料に書いてありますけども
何も輸入促進だけを言ってるわけではなくて
外国人がもっと入りやすくなってくると
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

本手法の出力：

ですからインポートって
本日のお手元の資料に書いてありますけども
何も輸入促進だけを言ってるわけではなくて
外国人がもっと入りやすくなってくると
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 2 正解データと本手法の出力結果の比較

4.2 実験結果

本手法及び従来手法[2]の適合率，再現率，F 値を表 2 にそれぞれ示す．本手法は，従来手法と比較して，適合率で2.55%下回ったが，再現率で4.88%，F値で0.96%，上回り，本手法の有効性を確認した．

一方，適合率及び再現率における分母に着目すると，正解や従来手法と比べ，本手法は改行を挿入する頻度が多く，本手法は余分に改行を挿入する傾向にあることがわかる．その傾向を示す例を図 2 に示す．本手法の出力は正解と比べ，頻繁に改行されており，結果として，短い行が多くなり，読みにくい字幕となっていることがわかる．

5 まとめ

本論文では，RNN を用いた講演音声のテキストへの改行挿入手法を提案した．比較実験の結果，F 値において従来手法[2]の 81.43%を上回る 82.39%を達成し，本手法の有効性を確認した．今後は，余分な改行を抑制するための素性の導入を検討したい．

謝辞 本研究は，一部，科学研究費補助金基盤研究(C) No. 16K00300 及び 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] 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，J98-D(6)，pp. 1621-1631，2009.
[3] S. Matsubara et al., “Bilingual Spoken Monologue Corpus for Simultaneous Machine Interpretation Research,” Proc. LREC 2002, pp. 153-159, 2002.
[4] M. Sundermeyer et al., “LSTM Neural Networks for Language Modeling,” Proc. INTERSPEECH 2012, pp. 194-197, 2012.