

漸進的な言語処理のためのRNNを用いた残存文長の推定とその評価 Estimation of Remaining Sentence Length for Incremental Language Processing by RNN and Its Evaluation

河村 天暉^{†,a)} 大野 誠寛^{†,b)} 松原 茂樹[†]
Takaki Kawamura Tomohiro Ohno Shigeki Matsubara

1 はじめに

同時通訳や字幕生成などの音声言語処理システムでは、入力に対して漸進的に処理する必要があり、処理の正確さを保ちつつ、遅延時間を抑えることが求められる。このような処理を実現するにあたり、意味的なまとまりをもつ文が今後どれだけ続くかという情報は、重要な手がかりとなりうる。例えば、文がもう少しで終わることが分かれば、処理の正確さを保つため、同時通訳における訳出タイミング [1] や、読みやすい字幕とするための改行挿入タイミング [2] を遅らせるという判断が可能となる。

これまでに著者らは、文節が入力されるごとに文の残りの長さ（以下、残存文長）を SVM を用いて推定する手法（以下、従来手法）を提案している [3]。この手法は、残存文長が 1 文節、2~3 文節、4 文節以上の 3 クラスのいずれであるかを推定するが、その精度は十分ではない。

そこで本稿では、推定精度の向上を目指し、RNN (Recurrent Neural Network) [4] を用いた残存文長の推定手法を提案する。また、従来手法や人間による推定結果との比較評価を行い、本手法の有効性を検証する。

2 残存文長

日本語文を構成する要素には、単語、文節、節などが存在するが、本研究では文の長さを測る単位として、文節を使用する。文 s が n_s 個の文節から成り、文頭から x 番目の文節 b_x まで既に入力されているとき（すなわち、既入力文節数が x であるとき）の残存文長 $RL(s, x)$ を $RL(s, x) = n_s - x$ により定義する。

3 RNN を用いた残存文長の推定

提案手法では、1 文 ($s = b_1 \dots b_{n_s}$) 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を RNN に入力し、そのときの残存文長（文節数）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-1} が入力されるまで繰り返す。

例えば、文 s 「私は / この間 / 学校に / 行った。」^{*1} の中の文節 b_2 「この間」が入力されたときの残存文長 $RL(s, 2)$ を推定する際には、RNN への入力は、文頭から現在入力された文節までの形態素系列「私はこの間」とし、RNN の出力は、そのときの残存文長 $RL(s, 2)$ の確率分布とする。この確率分布に基づいて、どの値を最終的な推定結果とするかについては様々な方法が考えられるが、提案手法では期待値を推定結果とする。

なお、RNN の出力層の次元数は、最も長いと想定する文の長さ（文節数）とする。また、RNN の隠れ層は 1 層かつ LSTM (Long Short Term Memory) [5] により構成され、RNN への各入力単語は one-hot ベクトルで表現されて

いるものとする。さらに、単語系列の入力方向については、文頭から順に入力していく順方向、文末からの逆方向、順方向と逆方向の出力を連結させた両方向の 3 種類を事前検証した結果 [6]、逆方向を採用することとした。

4 評価実験

従来手法 [3] との比較実験を実施し、残存文長の推定精度の向上における、提案手法の有効性を検証する。

従来手法は、残存文長が 1 文節、2~3 文節、4 文節以上の 3 クラスのいずれであるかを推定するものであるため、従来手法との精度比較を行うためには、提案手法の出力を、従来手法に合わせて 3 クラスに変換する必要がある。3 クラスに変換する方法は様々考えられるが [6]、本実験では、RNN が出力する確率分布の期待値（小数点第 1 位を四捨五入）が、変換先の 3 クラスの中でどのクラスに属するかを求め、そのクラスを提案手法の推定結果とする。

4.1 実験概要

実験データには、京都大学テキストコーパス Ver.4.0^{*2} のうち、文の長さが 1 文節であるものを除いた 38,115 文を使用した。このうち、1 月 4 日、1 月 5 日の全記事 2,291 文をテストデータ、1 月 9 日の全記事 1,212 文を開発データ、残りの 34,612 文を学習データとして用いる。形態素情報や文節境界情報は、当該コーパスのものをそのまま利用した。なお、音声言語処理システムへの応用を考えると、話し言葉特有の言語現象にも対処する必要があるが、本稿では、端緒として、整った文を対象とする。比較手法として、以下の 2 つの手法を用意した。

- ・ **ChanceRate** : 学習データより、各入力文節位置における 1 文の残存文長を集計し、その割合に従ってランダムに 3 クラスを出力する。

- ・ **従来手法 (SVM)** : 学習データの文集 S における入力文節数ごとの平均残存文長 ($ARL(S, x) = \frac{\sum_{s \in \{s | n_s > x, s \in S\}} RL(s, x)}{|\{s | n_s > x, s \in S\}|}$)、入力文節の語彙情報、既入力文節列の構文情報を素性とし、SVM により 3 クラスを出力する [3]。

評価では、クラスごとの適合率、再現率、F 値、および、全入力文節位置での推定結果に対する正解率（以下、文節単位正解率）、さらに、1 文中の全ての推定において正解した文の割合（以下、文単位正解率）を測定した。

RNN の学習は、Chainer V4.0^{*3} を介して行った。学習アルゴリズムには Adam を採用した。パラメータの更新はミニバッチ学習（学習率 0.001、バッチサイズ 128）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 50 とした。入力層の出力ベクトルと隠れ層の次元数は共に 200 とした。この値は、200, 400, 600, 800, 1000 の 5 通りで学習を行い、開発データの推定において最も正解率の高かったものを採用した。

単語の one-hot ベクトルの次元数は 29,725 とした。学習データ中の異なり語数 29,724 に未知語タグを加えたもの

[†] 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総括本部, Information and Communications, Nagoya University.

a) 18fmi07@ms.dendai.ac.jp

b) ohno@mail.dendai.ac.jp

*1 単語境界を半角スペース、文節境界をスラッシュで表す。

*2 <http://nlp.ist.i.kyoto-u.ac.jp/index.php> 京都大学テキストコーパス

*3 <https://chainer.org/>

表 1 3 クラス分類の実験結果 (適合率: P, 再現率: R, F 値: F, 文節単位正解率, 文単位正解率)

	残存文長: 1 文節			残存文長: 2~3 文節			残存文長: 4 文節以上			文節単位正解率 (%)	文単位正解率 (%)
	P(%)	R(%)	F	P(%)	R(%)	F	P(%)	R(%)	F		
Chance Rate	11.42 (269/2,420)	11.74 (269/2,291)	11.42	21.19 (978/4,615)	22.39 (978/4,368)	21.77	67.67 (9,173/13,556)	65.83 (9,173/13,934)	66.74	50.60 (10,420/20,593)	0 (0/2,291)
従来手法 (SVM)	26.66 (1,511/5,667)	65.95 (1,511/2,291)	37.97	30.36 (1,321/4,394)	30.24 (1,321/4,368)	30.15	79.17 (8,338/10,532)	59.84 (8,338/13,934)	68.16	54.25 (11,170/20,593)	0.79 (18/2,291)
提案手法 (RNN)	36.54 (768/2,102)	33.52 (768/2,291)	34.96	29.46 (1,605/5,448)	36.74 (1,605/4,368)	32.70	76.62 (9,994/13,043)	71.72 (9,994/13,934)	74.09	60.05 (12,367/20,593)	2.14 (49/2,291)



図 1 Web インタフェース画面

である。なお、テストデータにおける未知語率は 2.88% であった。一方、学習データにおける最長の文長は 47 文節であったため、RNN の出力層の次元数は 46 とした。

4.2 実験結果

表 1 に実験結果を示す。クラス「1 文節」の F 値は従来手法が最大値を得ているものの、他のクラスの F 値や文節単位および文単位正解率に着目すると、提案手法が従来手法を全体的に上回っていることがわかる。

以上から、3 クラス上での残存文長の推定における精度向上に提案手法が有効であることを確認した。

5 人間による残存文長推定との比較

4 節では、従来手法との比較により提案手法の有効性を確認したが、残存文長の推定タスクの難しさに対して、提案手法がどの程度の性能を示しているのかは検証できていない。そこで、人間が残存文長を推定する能力を測定し、残存文長の推定タスクの難しさの程度を確認するとともに、人間による推定結果と提案手法による自動推定結果を比較評価した。

5.1 人間の残存文長推定能力の測定

作業者は、1 文が文頭から順に 1 文節ずつ提示されるという状況において、新たな文節が提示されるごとに、これまでに提示された文節列を基に残存文長を推定する。具体的には、図 1 の Web インタフェースを用いて、次の手順により推定作業を実施した。

- 1) 左フレームより文 ID を選択し、その冒頭文節を表示。
- 2) 表示された既入力文節列を基に残存文長を推定し、入力欄に入力。
- 3) 「次へ」ボタンを押し、次の文節を表示。
- 4) 2) と 3) の推定・表示を文末まで繰り返す。

上述の作業を、対象データ 50 文に対して大学生 5 名がそれぞれ実施した。対象データは、4.1 節のテストデータの 1 月 4 日の全記事よりランダムに抽出した。

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

	F 値			文節単位正解率 (%)	文単位正解率 (%)
	1	2~3	4~		
人間 (平均)	43.30	35.07	65.35	54.24	5.60
人間 (最高)	43.75	37.23	74.95	61.68 (272/441)	10.00 (5/50)
提案手法	34.69	37.13	71.66	58.28 (256/441)	0.00 (0/50)

評価では、4 節と同様に比較するため、人間による残存文長の推定結果を、それぞれ 1 文節、2~3 文節、4 文節以上の 3 クラスのいずれかに分類した上で、クラスごとの F 値、文節単位および文単位の正解率を測定した。

5.2 比較評価

表 2 に比較評価結果を示す。人間 (平均) の値は 5 名のマクロ平均値、人間 (最高) の値は 5 名のうち、文節単位正解率が最高値だった作業者の値であり、提案手法の値は対象データ 50 文に対するものである。文節単位正解率に着目すると、提案手法は人間 (最高) を下回るものの、人間 (平均) を上回っていた。このことから、提案手法が人間と同程度の推定精度を達成していることを確認した。

一方、文単位正解率に着目すると、人間による推定では正解している文が存在しているのに対し、提案手法は 1 文も正解できていない。人間が捉えている特徴の中には、RNN が捉えきれていないものがあると考えられる。

6 おわりに

本論文では、RNN を用いて、1 文を構成する文節が入力されるごとに残存文長を推定する手法を提案した。比較実験の結果、従来手法を 5.80% 上回る文節単位正解率 60.05% を達成し、提案手法の有効性を確認した。今後は、Attention の導入などによる精度向上、また、話し言葉への対応に取り組みたい。

謝辞 本研究は、一部、科学研究費補助金基盤研究 (B) No. 26280082 及び (C) No. 16K00300 により実施した。

参考文献

- [1] 笠ら, “英日同時翻訳のための依存構造に基づく訳文生成手法,” 信学論, J92-D(6), pp. 921–933, 2009.
- [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, J92-D(9), pp. 1621–1631, 2009.
- [3] 河村ら, “漸進的な言語処理のための残存文長の推定,” FIT2018 講演論文集, 17(2), pp. 183–184, 2018.
- [4] Y. Bengio et al., “A Neural Probabilistic Language Model,” J. MACH. LEARN. RES, 3, pp. 1137–1155, 2003.
- [5] M. Sundermeyer et al., “LSTM Neural Networks for Language Modeling,” Proc. of INTERSPEECH 2012, pp. 194–197, 2012.
- [6] 河村ら, “漸進的な言語処理のための RNN を用いた残存文長の推定,” 言語処理学会第 25 回年次大会発表論文集, pp. 355–358, 2019.