P4-5 単語の難易度埋め込みを用いた日本語のテキスト平易化

柳本 大輝，梶原 智之，二宮 崇（愛媛大学）

1. 背景

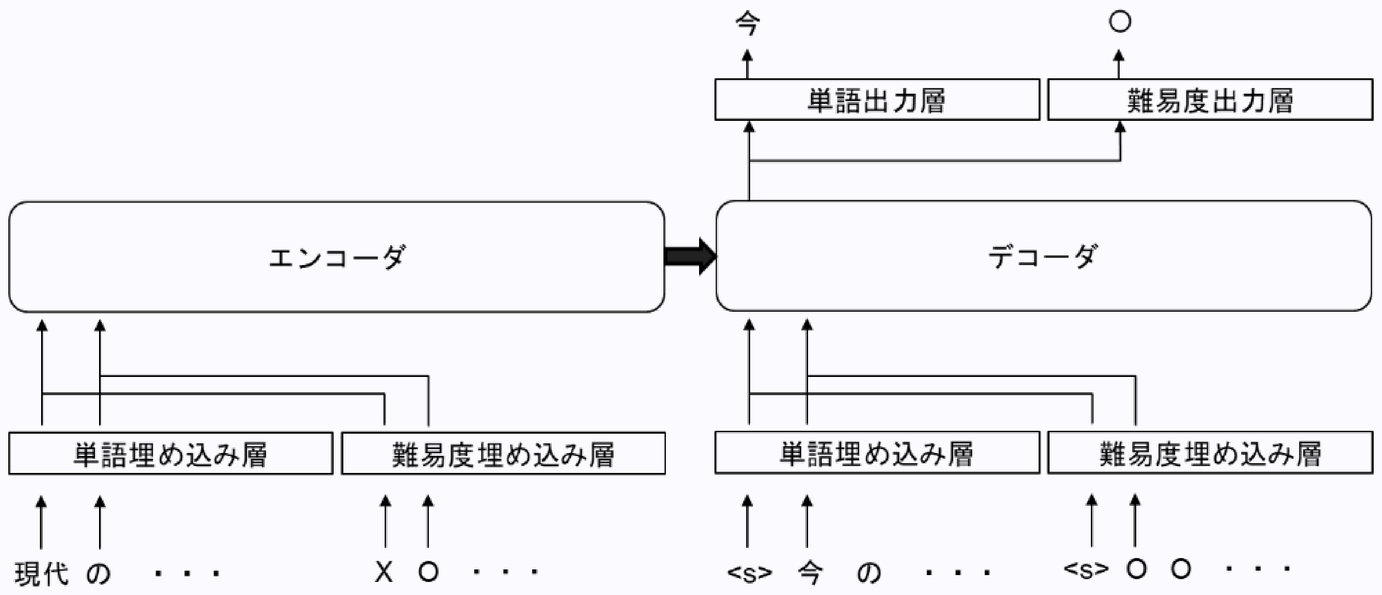
- 先行研究の多くは、テキスト平易化タスクの特徴である難易度を充分に考慮していない
- テキストの難易度を考慮する英語の先行研究は、多段階の文の難易度が付与されたパラレルコーパスを使用する
- 日本語には文の難易度が付与されたパラレルコーパスはないが、単語難易度辞書は存在する

2. 提案手法

- 単語の難易度を考慮したテキスト平易化モデルの訓練を行う
- Transformer に難易度埋め込み層を追加し、単語が基礎語彙に含まれるか否かで2種類（O, X）の特殊トークンを難易度埋め込み層に入力する
- 単語 x_i 、難易度 g_i をそれぞれの埋め込み層に入力し、生成された埋め込みを連結する

$$Emb = Wx_i \parallel Fg_i$$

- 単語および難易度は、エンコーダとデコーダで共通のため、どちらも埋め込み層を共有する



3. 実験設定

- **データセット**
  - やさしい日本語コーパス SNOW
  - SNOWに付属する基礎語彙 2,000 単語

文対数		
訓練	検証	評価
82,300	2,000	100×7

- **比較手法**
  - ベースライン　：単語難易度を考慮しないTransformer（6 層 8 ヘッド 512 次元）
  - 出力の語彙制限　：デコーダの語彙を基礎語彙 2,000 単語に制限
  - 難易度考慮　：単語難易度を考慮する提案手法
  - 難易度考慮（○○を非共有）：○○埋め込み層を共有しない

4. 実験結果

モデル	BLEU	SARI	add	keep	del	基礎語彙の割合
ベースライン	75.55	63.88	16.78	88.23	86.63	77.79
出力の語彙制限	43.75	50.20	12.19	63.45	74.96	**100.00**
難易度考慮	74.25	64.10	**17.56**	88.00	86.75	78.03
難易度考慮（単語層・難易度層を非共有）	50.17	54.20	11.83	71.40	79.37	79.90
難易度考慮（単語層を非共有）	74.37	**64.32**	17.47	**88.35**	**87.16**	77.34
難易度考慮（難易度層を非共有）	**75.59**	64.00	17.10	88.19	86.70	78.14

- 単語難易度を考慮することで、BLEU、SARI が向上した
- 単語層のみを共有しないときに SARI が最も高い
  → 語彙が多いエンコーダ側（約 20,300）に対して、語彙が少ないデコーダ側（約 5, 000）では、ひとつひとつの単語が広い意味を持つ必要があるため、単語埋め込みは共有しない方が良い可能性がある

入力文	**彼 の** 援助 **の おかげ で 、 私 の 仕事 は** 現在 順調 **に 進ん で ます 。**
ベースライン	**彼 が 助け て くれ た おかげ で 、 私 の 仕事 は 今日 調子 よく 進ん で ます 。**
出力の語彙制限	**彼 が** <unk> **て くれる おかげ で** <unk> **私 の 仕事 は 今 調子 が 悪い** <unk>
難易度考慮	**彼 の 協力 の おかげ で 、 私 の 仕事 は 予定 通り に 進ん で ます 。**
難易度考慮（単語層・難易度層を非共有）	**彼 が 助け て くれ た おかげ で 、 私 の 仕事 は 今 予定 通り です 。**
難易度考慮（単語層を非共有）	**彼 が 助け て くれる から 、 私 の 仕事 は 今 、 進ん で 行き ます 。**
難易度考慮（難易度層を非共有）	**彼 の 協力 の おかげ で 、 私 の 仕事 は 今 早く 進ん で ます 。**
参照文	**彼 が 助け て くれ た の で 、 私 の 仕事 は 今 予定 通り 進ん で ます 。**