編集距離に基づくカリキュラム学習を用いたスタイル変換

門谷宙* 梶原智之** 荒瀬由紀* 鬼塚真*

*大阪大学大学院情報科学研究科 **愛媛大学大学院理工学研究科

OSAKA UNIVERSITY

もくじ

[1] Platanios et al. (NAACL 19) Competence-based Curriculum Learning for Neural Machine Translation

もくじ

[1] Platanios et al. (NAACL 19) Competence-based Curriculum Learning for Neural Machine Translation

スタイル変換

- 入力文の意味を保持したまま表現を変更するタスク
- 単言語パラレルコーパス上で機械翻訳と同様の手法を使用
- 応用事例: 文章読解支援, 機械翻訳の前処理

カリキュラム学習

- 簡単な問題から学習を始め, 徐々に難しい問題を学習
- 機械学習モデルの性能向上
- 適用事例: 物体認識 [2], マルチメディア検索 [3]

[2] Xiao et al. (ACMMM 19) Error-Driven Incremental Learning in Deep Convolution Neural Network for Large-Scale Image classification
[3] Jiang et al. (NIPS 14) Self-Paced Learning with Diversity

自然言語処理におけるカリキュラム学習

- **機械翻訳**におけるカリキュラム学習
  - 訓練サンプルの難易度のみを考慮する手法 [4]
    → 学習の収束は早くなるが, 翻訳品質は向上せず
  - モデルの能力を考慮する手法 [1]
    → 収束後においても翻訳品質が向上
- **スタイル変換**におけるカリキュラム学習の先行研究はなし

[4] Zhang et al. (arXiv 18) An Empirical Exploration of Curriculum Learning for Neural Machine Translation

もくじ

[1] Platanios et al. (NAACL 19) Competence-based Curriculum Learning for Neural Machine Translation

Platanios et al. [1]: 手法概要

- 機械翻訳におけるカリキュラム学習手法
- 2つの指標を導入
  - 訓練サンプル s_i の難易度$\bar{d}(s_i) \in [0, 1]$
  - 訓練ステップ tにおける モデルの能力$c(t) \in [0, 1]$
- 各ステップで $\bar{d}(s_i) \leq c(t)$を満たす訓練サンプルのみを使用
  → 訓練時間の経過に伴って使用できる訓練サンプルが増加

Platanios et al. [1]: 難易度の基準$d(s_i)$

- 訓練サンプル s_iの入力文は単語列$\{w_1, \cdots, w_{N_i}\}$で構成される
- 難易度の基準$d_{length}(s_i)$
  - 長文は難しい → **文長**が難易度の指標
  - $d_{length}(s_i) \triangleq N_i$
- 難易度の基準$d_{rarity}(s_i)$
  - 低頻度語は難しい → **単語の出現頻度**が難易度の指標
  - $d_{rarity}(s_i) \triangleq -\sum_{j=1}^{N_i} \log \hat{p}(w_j)$　($\hat{p}(w_j)$: 単語w_jの出現確率)
- 既存の難易度の基準は, 正解文を考慮していない

もくじ

[1] Platanios et al. (NAACL 19) Competence-based Curriculum Learning for Neural Machine Translation

スタイル変換における難易度

- ほとんど変換を必要としない訓練サンプル:
  入力文をコピーするだけで, 正解文とほぼ一致 (簡単)
- 多くの変換が必要な訓練サンプル:
  複雑な書き換え操作が必要 (難しい)
- 入力文を正解文に変換するために必要な変換コストと仮定
  → カリキュラム学習に**編集距離**を導入

入力文	正解文
Their first two albums were **pretty** good.	Their first two albums were **very** good.
no where there is no such thing	**That does not exitst.**

編集距離

- 単語列Xを単語列Yに変換するために必要な編集操作の回数
  - 挿入　単語を1つ加える
  - 削除　単語を1つ消す
  - 置換　単語を1つ別の単語に変える
- 簡単な訓練サンプルは小さく, 難しい訓練サンプルは大きい

入力文	正解文	編集距離
Their first two albums were **pretty** good.	Their first two albums were **very** good.	**1**
no where there is no such thing	**That does not exitst.**	**7**

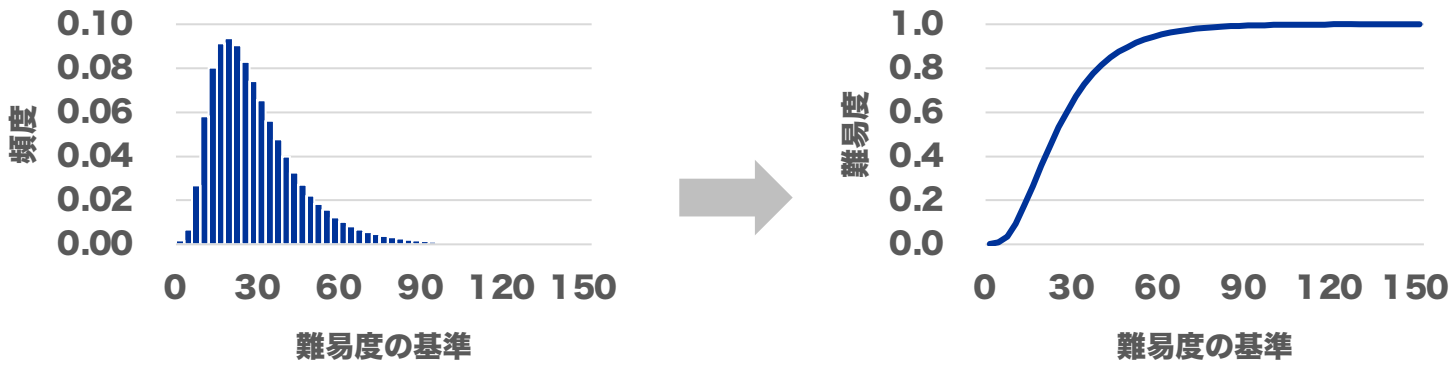
提案手法: 編集距離に基づくカリキュラム学習

- スタイル変換に初めてカリキュラム学習を適用
- カリキュラム学習の枠組みは Platanios et al. [1] に従う
- 難易度の基準 $d_{distance}(s_i)$
  - 難易度の指標: 編集距離
  - $d_{distance}(s_i) \triangleq E_i$
    E_i: 訓練サンプル s_iの入力文と正解文の編集距離
  - 入力文と正解文の両方を考慮

カリキュラム学習の枠組み: Platanios et al. [1]

難易度$\bar{d}(s_i)$

- 難易度の基準$d(s_i)$から難易度$\bar{d}(s_i)$への変換手順
  1. $d(s_i)$に従って累積分布関数を作成
  2. 累積分布関数上で$d(s_i)$に対応する値を$\bar{d}(s_i)$とする
- $\bar{d}(s_i)$が小さいほど簡単, 大きいほど難しい



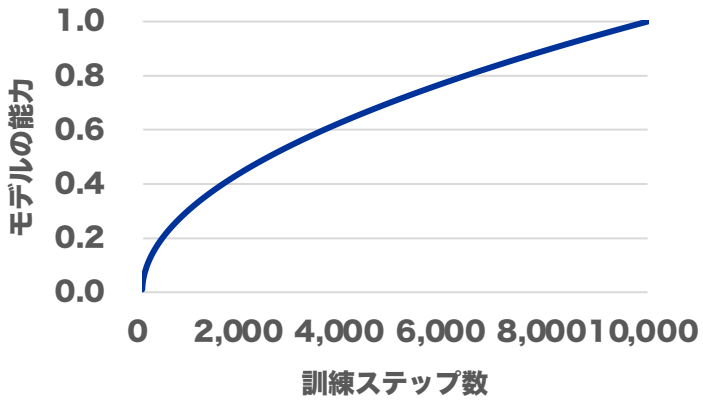
カリキュラム学習の枠組み: Platanios et al. [1]

モデルの能力 $c(t)$

- $c(t) = \min(1, \sqrt{t\frac{1-{c_0}^2}{T} + {c_0}^2})$

  c_0: モデルの能力の初期値
  T: モデルの能力が完全に備わると予想されるステップ数

- $c(t)$は訓練開始時は小さく, 訓練の経過に伴い単調増加



もくじ

[1] Platanios et al. (NAACL 19) Competence-based Curriculum Learning for Neural Machine Translation

実験設定

フォーマルさに関するスタイル変換の性能を評価

- データセット: GYAFC [5]
- スタイル変換モデル: Transformer [6]
- 評価指標: BLEU [7]

	Train	Dev	Test
E&M	209,124	2,877	1,416
F&R	209,124	2,788	1,332

[5] Rao and Tetreault (NAACL 18) Dear Sir or Madam, May I Introduce the GYAFC Dataset: Corpus, Benchmarks and Metrics for Formality Style Transfer
[6] Vaswani et al. (NIPS 17) Attention is All you Need
[7] Papineni et al. (ACL 02) Bleu: a Method for Automatic Evaluation of Machine Translation

比較手法

- ベースライン　　カリキュラム学習を用いない手法
- CL-SL　　**文の長さ**に基づくカリキュラム学習
- CL-SR　　**単語の出現頻度**に基づくカリキュラム学習
- CL-ED　　**編集距離**に基づくカリキュラム学習

実験結果

- 両ドメインで, 提案手法がベースラインを上回る性能を達成
- 既存のカリキュラム学習は有効でないが, 提案手法は有効

	カジュアル → フォーマル	
	E&M	F&R
入力文	49.19	50.94
正解文	100.0	100.0
ベースライン	69.81	75.02
CL-SL	69.83	74.90
CL-SR	70.05	74.62
CL-ED	**70.34**	**75.41**

もくじ

[1] Platanios et al. (NAACL 19) Competence-based Curriculum Learning for Neural Machine Translation

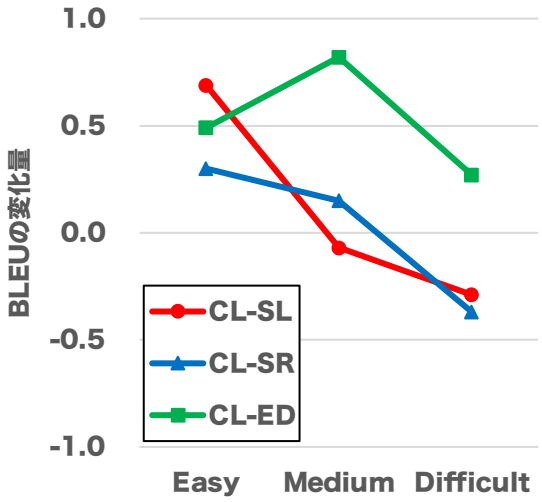
分析手順

どのような特性を持つ事例に対する性能が向上するか分析

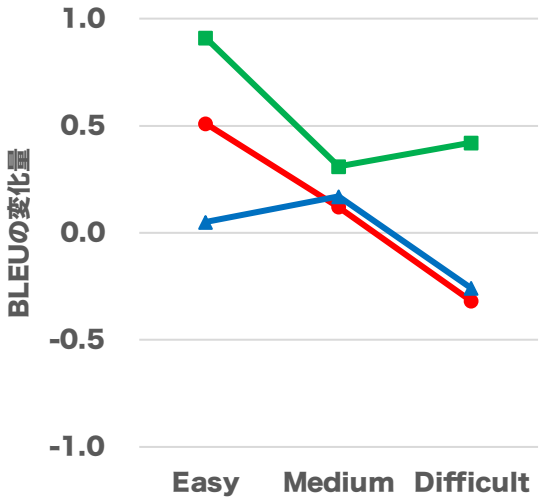
- 評価データを難易度の指標に従って, 3つのビンに振り分け
- 比較手法を用いて, ビン毎にBLEUを測定
  - ベースライン　カリキュラム学習を用いない手法
  - CL-SL　　**文の長さ**に基づくカリキュラム学習
  - CL-SR　　**単語の出現頻度**に基づくカリキュラム学習
  - CL-ED　　**編集距離**に基づくカリキュラム学習
- ベースラインからのBLUEの変化量を調べる

分析結果

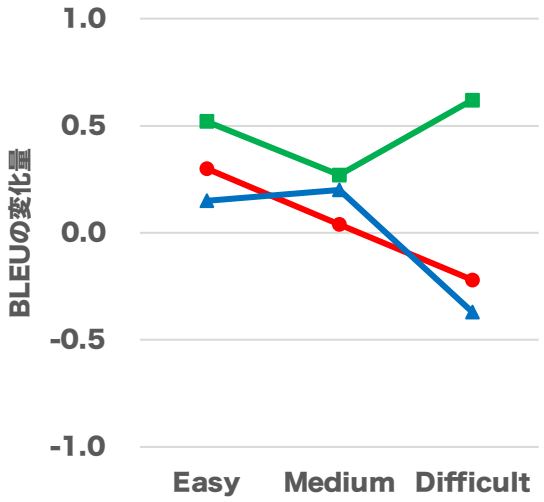
- 全体的に簡単な事例に対する性能の向上が大きい
- 既存のカリキュラム学習は難しい事例に対する性能が悪化
- 提案手法は難しい事例に対する性能を改善



(a) 文長



(b) 単語の出現頻度



(c) 編集距離

もくじ

[1] Platanios et al. (NAACL 19) Competence-based Curriculum Learning for Neural Machine Translation

まとめ

- 提案手法: 編集距離に基づくカリキュラム学習
  - スタイル変換に初めてカリキュラム学習を適用
  - 難易度の指標として編集距離を導入
- 評価実験の結果, 提案手法の有効性を確認
- 提案手法は難しい事例に対する性能改善に貢献