目的言語の低頻度語の高頻度語への言い換えによるニューラル機械翻訳の改善

首都大学東京　関沢祐樹　梶原智之　小町守　sekizawa-yuuki@ed.tmu.ac.jp

背景と目的

背景:
ニューラル機械翻訳:語彙次元の分類問題
計算量を削減するため語彙制限
低頻度語は特殊記号で出力される
→ 妥当性、流暢性が失われる

目的:
出力文の低頻度語(OOV)を削減する

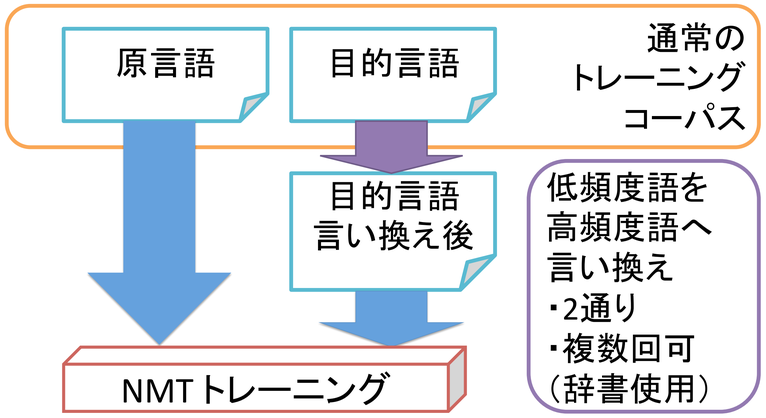
先行研究

- 訓練方法の変更によって出力文の OOV を削減
  - Jean et al. (2015)
  - Mi et al. (2016)
  - Luong et al. (2016)
    →訓練方法を変更する必要がある

- 前処理・後処理によって出力文の OOV を削減
  - Luong et al. (2015)
    →翻訳文対のアライメントが必要
  - Sennrich et al. (2016)
    →意味を考慮せず貪欲な手法

提案手法

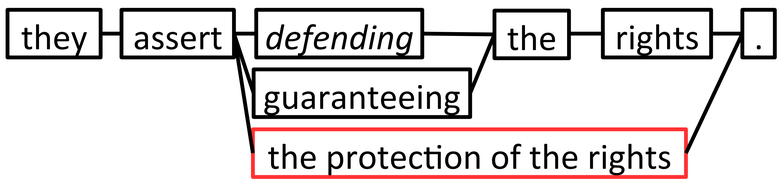
トレーニングデータのうち
目的言語の低頻度語を
同義な高頻度語に言い換え



言い換え1:言い換え確率を最大化する言い換え
妥当性を優先

原文	:the *pedagogues* had *quarrels*.
1回目の言い換え	:the *educators* had *discussions*.
2回目の言い換え	:the *teachers* had *discussions*.

言い換え2:言語モデル確率が最大化する言い換え
流暢性を優先



実験

日英翻訳結果
BLEUスコア(出力文に存在する OOV の数)

手法		言い換え確率	言語モデル確率
Bahdanau+ (2015)		20.63 (1,489)	
1 回のみ	(語)	20.55 (1,240)	20.49 (1,338)
2 回まで	(語)	20.61 (1,301)	**20.71** (1,231)
無制限	(語)	20.28 (1,322)	18.23 (1,229)
1 回のみ	(語＋句)	20.11 (1,274)	17.89 (1,451)
2 回まで	(語＋句)	19.29 (1,408)	18.38 (1,442)
無制限	(語＋句)	19.61 (1,324)	18.65 (1,327)

設定
- コーパス:アジア学術論文抜粋コーパス(ASPEC)
  トレーニング:827,503文対
- 言い換え辞書:PPDB
- 言語モデル:ASPEC　2-gram
- 翻訳モデル:Bahdanau et al. (2015)
- 語彙数:入出力ともに30,000語

翻訳例

改善例

手法	翻訳
input	オゾン 生成 量 が 約 2 mg / h 増大 し た 。
reference	ozone formation increased about 2mg / h .
Bahdanau+ (2015)	the amount of ozone generation increased by about “OOV” / h .
2 回まで (語、言語モデル確率)	the ozone generation increased by about 2 mg / h .

改悪例

手法	翻訳
input	後者 の コイル は 液体 ヘリウム 中 で 2 . 2 t を 出し た 。
reference	the latter coil generated 2.2t in liquid helium .
Bahdanau+ (2015)	the latter coil was 2.2 t. in the liquid helium .
2 回まで (語、言語モデル確率)	the latter coil was “OOV” in liquid helium .

考察

- 言い換えは1回よりも2回の場合にBLEUスコアが高い
- 無制限に言い換え繰り返し → さらなる改善はなし
  言い換えの意味の保持と OOV 削減のバランスを取ると良い
- 語のみの言い換えは(語＋句)の言い換えよりBLEU スコアが高い
  句の内部の言語モデル確率を考慮していないため
  流暢性を損なう言い換えが行われた可能性がある
- トレーニングデータの OOV の減少に伴って
  翻訳結果の OOV が削減されるとは限らない
  出力文の流暢性の担保のために言い換え結果を出力しない