# P02　クロスリンガルな単語分散表現を用いた機械翻訳自動評価手法の検討

首都大学東京　小町研究室　　嶋中宏希　山岸駿秀　松村雪桜　小町守

## 研究の背景と概要

従来の自動表手法で最も広く用いられている手法であるBLEUでは出力文と参照文の表層の比較をするので、下のような例文において意味を考慮した評価ができない。そこで、先行研究 [Wang et al., HyTra 2016] では BLEU をモノリンガルな単語分散表現を用いる手法に変更するにより、意味を考慮した評価ができるという自動評価手法を提案している。また、クロスリンガルな単語分散表現を用いることにより Crosslingual Document Classification や Word Similarity Evaluation の精度を向上させたという先行研究が多く存在する。

そこで、本研究では [Wang et al., HyTra 2016] の手法をクロスリンガルな単語分散表現を用いるモデルに変更することにより、モノリンガルなものを用いるモデルより人手との相関が高くなるのではないかという考えのもと検討を行った。

<例>　出力文 : He is a nice player　←表層ではなく単語分散表現を比較→　参照文 : He is a great golfer

## 従来手法 (BLEU_modif [Wang et al., 2016])

$T$ = 出力文　$R$ = 参照文　$ng = ngram$　$\gamma$ = しきい値

$$\mathrm{BLEU}_{\mathrm{modif}} = BP(T,\mathrm{R}) \cdot \exp\left(\sum_{n=1}^{4} \frac{1}{4} \log P_{sim}\right)$$

$$P_{sim} = \sum_{ng \in T} \frac{Max\_sim_{pruned}(ng,T,R,\gamma)}{\sum_{ng \in T} Count(ng)}$$

$$BP(T,R) = \begin{cases} 1 & \dots\ if\ len(T) > len(R) \\ e^{1-\frac{len(T)}{len(R)}} & \dots\ else \end{cases}$$

$Max\_sim_{pruned}(ng,T,R,\gamma)$ は 出力文 $T$ の $ngram$ 平均ベクトルと参照文 $R$ の全ての $ngram$ 平均ベクトル との cos 類似度の最大値であり、その最大値がしきい値以下の場合 0とする

## 提案手法 (BLEU_modif_C)

1. MultiVec [Berard et al., LREC 2016] の bivec を用いてクロスリンガルな単語分散表現を学習。Uniform Alignment を使用。

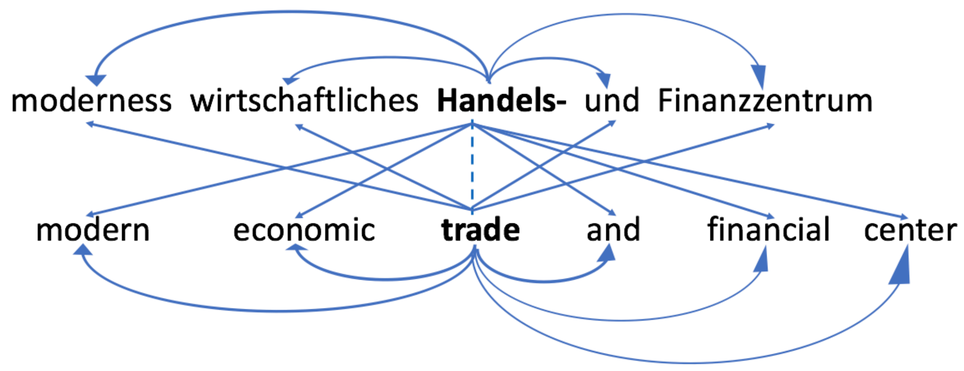


図 1　bivec (SkipGram) の簡易図

2. BLEU_modif の手法の出力文と参照文の類似度の計算を行う部分を1で作成したクロスリンガルな単語分散表現を用いて出力文と原文で類似度の計算を行いスコアを計算

## 実験設定

- Train (モノリンガルな単語分散表現)
  - Gensim の Word2Vec [Mikolov et al., 2013] を用いて学習
    SkipGram (dimension : 300, window_size : 5, epoch : 10)
    CBOW (dimension : 300, window_size : 5, epoch : 20)
- Train (クロスリンガルな単語分散表現)
  - MultiVec [Berard et al., LREC 2016] の bivec モデルを用いて学習
    SkipGram (dimension : 300, window_size : 5, epoch : 10)
    CBOW (dimension : 300, window_size : 5, epoch : 20)
  - Data (Parallel) : Europarl-v7 (WMT16) (En : 約 4,119 万語, De : 約 3,928 万語) (モノリンガルの単語分散表現も同じものを用いた)
- Dev
  WMT14 metrics task (segment level) (En → De : 2,737文, 18 システム)  (De → En : 3,003文, 13 システム)
- Test
  WMT15 metrics task (segment level) (En → De : 2,169文, 16 システム)  (De → En : 2,169文, 13 システム)
- 人手評価 (評価者 : 5人, 評価 : 5段階)
  ケンドールの順位相関係数を使用

## 実験結果

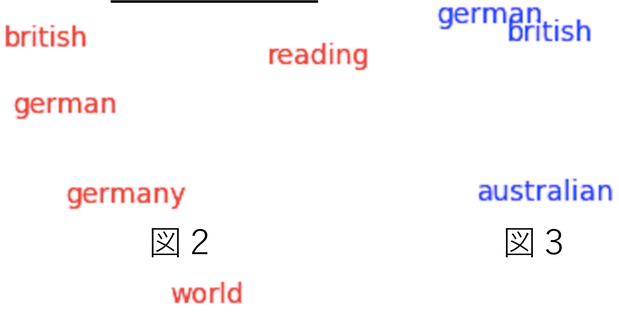


図 2




図 3


図 4




図 5

図 2～5 クロスリンガルな単語分散表現 (En : CBOW) (赤) と モノリンガルな単語分散表現 (En : CBOW) (青)

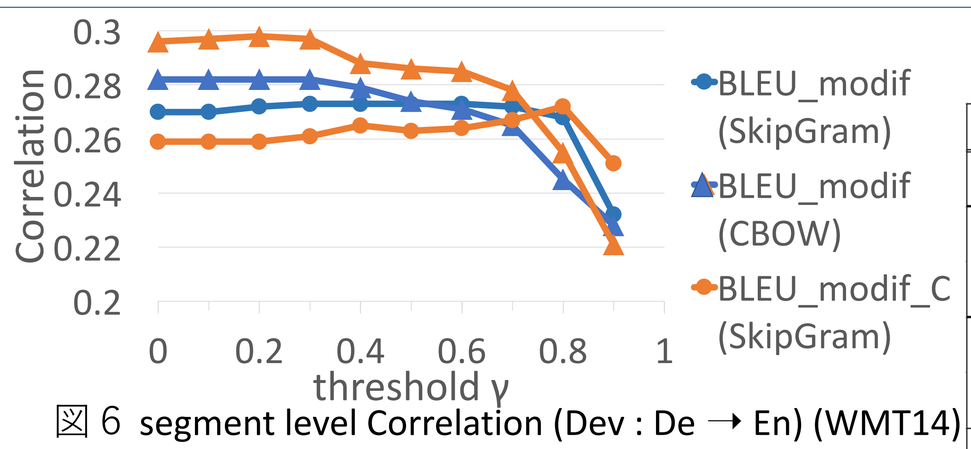


図 6　segment level Correlation (Dev : De → En) (WMT14)

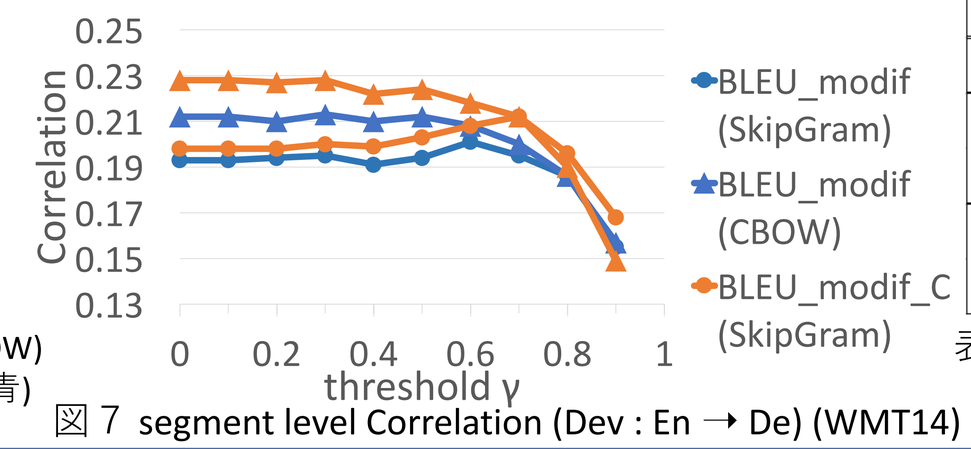


図 7　segment level Correlation (Dev : En → De) (WMT14)

| En → De | | |
|---|---|---|
| sent_BLEU | | 0.293 |
| BLEU_modif | SkipGram | 0.313 ($\gamma$ = 0.6) |
| | CBOW | 0.313 ($\gamma$ = 0.3) |
| BLEU_modif_C | SkipGram | 0.306 ($\gamma$ = 0.7) |
| | CBOW | **0.337** ($\gamma$ = 0.3) |
| De → En | | |
| sent_BLEU | | 0.360 |
| BLEU_modif | SkipGram | 0.357 ($\gamma$ = 0.5) |
| | CBOW | 0.376 ($\gamma$ = 0.0) |
| BLEU_modif_C | SkipGram | 0.334 ($\gamma$ = 0.8) |
| | CBOW | **0.391** ($\gamma$ = 0.2) |

表 1　segment level Correlation (Test) (WMT15)

## 考察

- クロスリンガルな単語分散表現のほうがよりトピック的に意味が似ている語が似たような分散表現で表されている（図 2, 図 3）
- クロスリンガルな単語分散表現のほうが原型が同じ語がほとんど同じ分散表現で表されていたり、意味が似ている語が似たような分散表現で表されている（図 4, 図 5）

→ 評価に単語分散表現を用いる場合、品詞的に似ている語よりトピック的に似ている語を似たような分散表現で表すほうが、人手との相関が高くなると考えられる

→ 今後、他の単語分散表現を用いる評価手法や、様々なクロスリンガルな単語分散表現の学習法で実験を行う必要がある