

# BERT のパラメータを用いた 最適輸送距離に基づく言い換え識別の性能評価

IBIS2021

山際宏明<sup>1</sup> 横井祥<sup>2,3</sup> 下平英寿<sup>1,3</sup>

<sup>1</sup> 京都大学大学院 <sup>2</sup> 東北大学大学院 <sup>3</sup> 理化学研究所革新知能総合研究センター

2021年11月10日

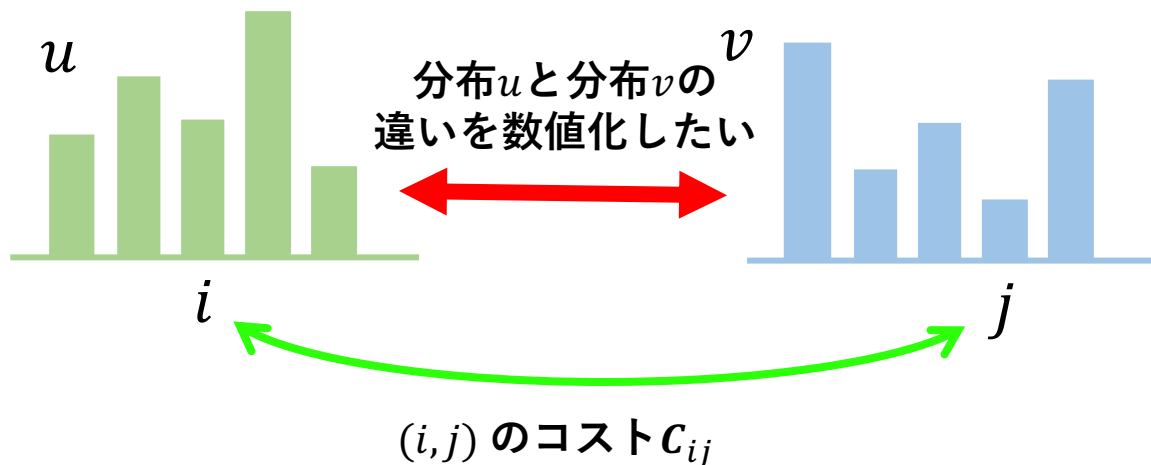
## 概要

最適輸送に基づく方法は共通語が含まれていない場合でも文の類似度を計算できるが、文中の語順を考慮することができない。そこでBERTが語順をencodeする構造を持つ事に着目し、事前学習済みのBERTに文を入力した際のパラメータを用いた最適輸送距離を考える。同じ単語で異なる意味となる文を含むデータセットを用いて、提案手法と最適輸送距離を用いた既存手法との性能を比較し、結果について考察する。

# 最適輸送距離

## モチベーション

- 分布  $u, v$  の距離の測定



## 設定

- 要素の組ごとにコストを設定
- 「コスト × 対応の割合」の総和を最小化

## 定式化

$$\text{最適輸送距離 } L_C(u, v) := \min_{P \in \Pi(u, v)} \sum_{i, j} c_{ij} P_{ij} \quad (1)$$

コスト  $c_{ij}$  が距離の時,  
**Wasserstein距離** という。

実現可能な  
対応の集合

全ての組について  
対応を考える

# Wasserstein 距離を用いた文類似度の計算

## 特徴

- 文 = 単語ベクトルの集合
- 共通語の無い文でも計算可能

## 既存手法:

**Word Mover's Distance (WMD)**  
(Kusner et al., 2015)

- 一様分布
- $C_{ij} = \|w_i - w_j\|_2$

## 類似手法:

**Word Rotator's Distance (WRD)**  
(Yokoi et al., 2020)

- $\|w_i\|$  で重み付けした分布
- $C_{ij} = 1 - \cos(w_i, w_j)$

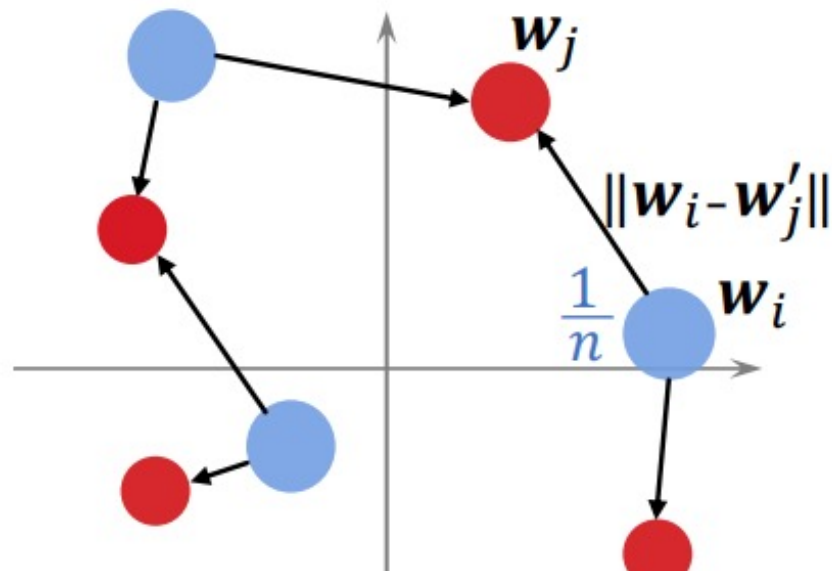


図1: (Yokoi et al., 2020) figure2 より引用

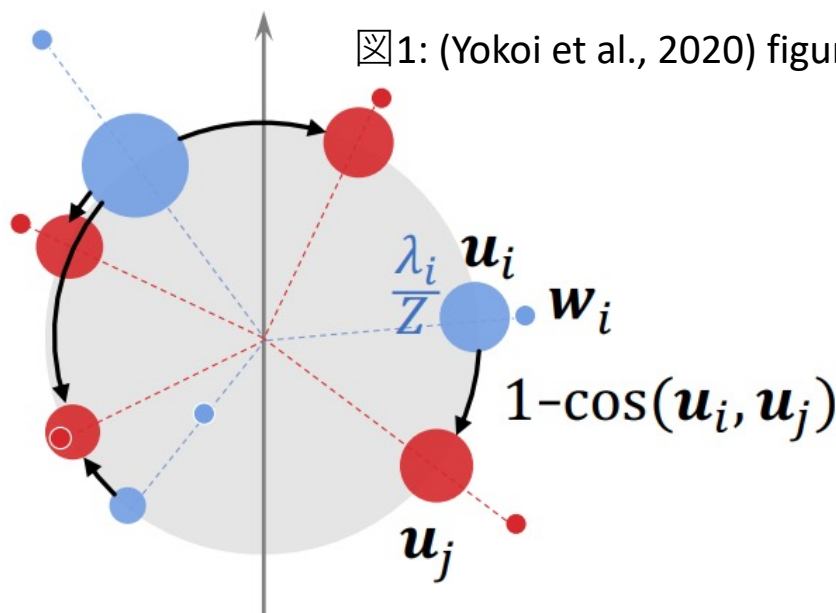
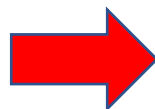


図2: (Yokoi et al., 2020) figure5 より引用

## 文間の単語の対応だけを考えて最適輸送距離

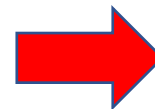
文の語順を考慮していない

 語順によって意味が大きく変わる文に対処できない

例

• Flights from **New York** to **Florida**

• Flights from **Florida** to **New York**



WMDの値は0.  
意味は**真逆**

## モチベーション

「文間の単語の対応 + 各文内の単語による構造情報」となる距離の提案

Wasserstein距離

Self-Attention &  
Gromov-Wasserstein距離

### 道具 1: Self-Attention

Transformer の Self-Attention は**構造情報**に関連 (Vaswani et al., 2017)

構造情報を持つ行列として  
Self-Attentionを利用

### 道具 2: Gromov-Wasserstein 距離

Gromov-Wasserstein 距離は**構造**の類似度を測定.

## 準備

- 入力文:  $\mathbf{X} \in \mathbb{R}^{n \times d}$
- Query:  $\mathbf{Q} := \mathbf{X}\mathbf{W}^Q \in \mathbb{R}^{n \times d_k}$
- Key:  $\mathbf{K} := \mathbf{X}\mathbf{W}^K \in \mathbb{R}^{n \times d_k}$

## Self-Attention

モデル内のパラメータ

$$\text{Self-Attention}(\mathbf{Q}, \mathbf{K}) := \text{softmax} \left( \frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}} \right) \in \mathbb{R}^{n \times n} \quad (2)$$

## Multi-head Attention

$0 < y < 1$  の値  スコアとみなせる

Transformer (Vaswani et al., 2017), BERT (Devlin et al., 2019) などのモデル

- 各 layer について複数の Self-Attention が存在
- それぞれ異なる特徴を捉える役割

# Self-Attention

Self-Attention の各成分が持つ情報 (Vaswani et al., 2017)

- **構造情報**
- 各文内の単語と単語の結びつきの強さ

pre-train された BERT に文を入力  
➡ Self-Attention を取り出せる

例 a woman is dancing.

低 (woman, a) = 0.01

高 (woman, dancing) = 0.8

BERTのpre-train時に  
用いるトークン

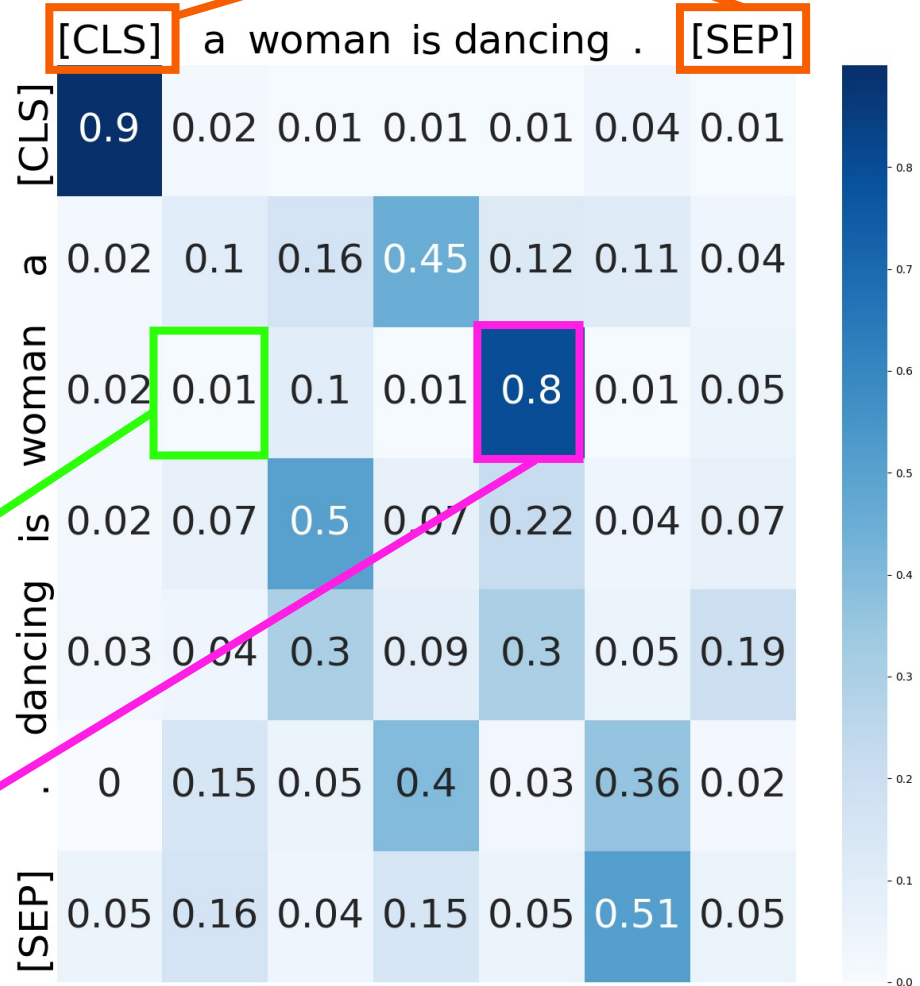


図3: “a woman is dancing.”を pre-train されたBERT に入力した際の Self-Attention

# Gromov–Wasserstein 距離

## 準備

$n$  個の要素と  $m$  個の要素について, 集合内でのコスト行列:  $D \in \mathbb{R}^{n \times n}$ ,  $D' \in \mathbb{R}^{m \times m}$

## Gromov–Wasserstein 距離 (Mémoli, 2011)

$$GW(\mathbf{u}, \mathbf{v}) := \min_{P \in \Pi(\mathbf{u}, \mathbf{v})} \sum_{i, j, i', j'} |D_{ii'} - D'_{jj'}| P_{ij} P_{i'j'} \quad (3)$$

$i, i'$  間のコスト       $j, j'$  間のコスト

$i \mapsto j, i' \mapsto j'$  と対応させた時,  
どれくらい2点間の距離(コスト)が変化するか

をすべての組同士で総和

## Fused Gromov–Wasserstein 距離 (Vayer et al., 2018, 2019) ( $\alpha \in [0, 1]$ )

$$FGW(\mathbf{u}, \mathbf{v}) := \min_{P \in \Pi(\mathbf{u}, \mathbf{v})} \sum_{i, j, i', j'} \left\{ (1 - \alpha) C_{ij} + \alpha |D_{ii'} - D'_{jj'}| \right\} P_{ij} P_{i'j'} \quad (4)$$

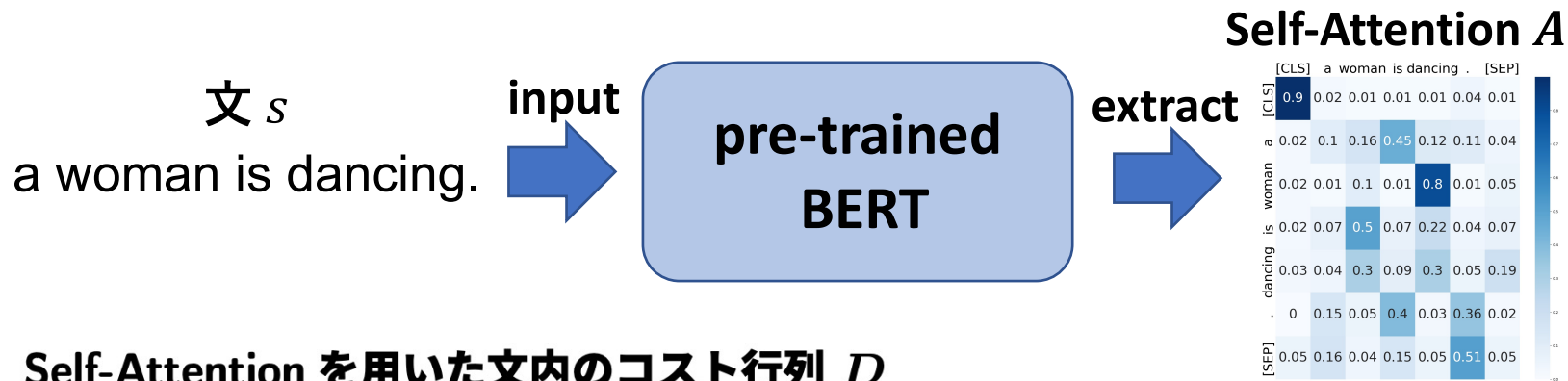
Wasserstein 距離

Gromov–Wasserstein 距離



## 準備

pre-train された BERT に文  $s$  を入力し Self-Attention  $A \in \mathbb{R}^{n \times n}$  を取り出す。



## Self-Attention を用いた文内のコスト行列 $D$

- Self-Attention の成分  $A_{ii'} \in [0, 1]$  は  $w_i, w_{i'}$  についての関連の度合い
- 文内の  $w_i, w_{i'}$  間のコスト =  $1 - A_{ii'}$ .  $0 < y < 1$  の値
- $I_n \in \mathbb{R}^{n \times n}$  を成分が全て 1 の行列として,  $D = I_n - A$ .

## アイデア

Wasserstein 距離 + コスト行列  $D$  を用いた Gromov–Wasserstein 距離

文内の単語の構造情報

## 定式化

Wasserstein 距離に用いる分散表現によって以下の2つを提案.

- BERT の分散表現を用いる場合

$$d_\alpha(\mathbf{u}, \mathbf{v}) := \min_{P \in \Pi(\mathbf{u}, \mathbf{v})} \sum_{i,j,i',j'} \left\{ (1 - \alpha) \underbrace{C_{ij}}_{C \in \mathbb{R}^{n \times m}} + \alpha \left| \underbrace{D_{ii'}}_{D \in \mathbb{R}^{n \times n}} - \underbrace{D'_{jj'}}_{D' \in \mathbb{R}^{m \times m}} \right| \right\} P_{ij} P_{i'j'} \quad (5)$$

- その他の分散表現を用いる場合

$$d_\alpha(\mathbf{u}, \mathbf{v}) := (1 - \alpha) \min_{P \in \Pi(\mathbf{u}, \mathbf{v})} \sum_{i,j} \underbrace{C_{ij}}_{C \in \mathbb{R}^{n' \times m'}} P_{ij} + \alpha \min_{Q \in \Pi(\mathbf{u}, \mathbf{v})} \sum_{i,j,i',j'} \left| \underbrace{D_{ii'}}_{D \in \mathbb{R}^{n \times n}} - \underbrace{D'_{jj'}}_{D' \in \mathbb{R}^{m \times m}} \right| Q_{ij} Q_{i'j'} \quad (6)$$

異なる tokenizer



$n \neq n', m \neq m'$

## WMD との比較

語順により意味が異なる文の組で性能比較

## PAWS データセット (Zhang et al., 2019)

共通語の多い文の組で言い換え識別を行うデータセット



# 実験結果

WMD と提案手法について PAWS の test (8000 組) を用いて AUC で比較。  
Self-Attention を取り出す BERT の layer, head 及び  $\alpha$  は dev (8000 組) での性能で決めた。  
複数の Self-Attention を選んだ場合は最適輸送距離を平均して最終的なスコアとした。

Table 1: BERT の 0 層目の embedding, word2vec, glove, fasttext を WMD に用いた時の AUC

Models	BERT-layer0	word2vec	glove	fasttext
WMD	0.5732	0.4897	0.4910	0.4928
ours	<b>0.6546</b>	<b>0.5164</b>	<b>0.4978</b>	<b>0.4986</b>


Table 2: test の評価時に用いた, BERT の layer, head,  $\alpha$  の組

	BERT-layer0	word2vec	glove	fasttext
	1/4/0.99	2/8/0.98	2/8/0.96	2/8/0.97
layer/head/ $\alpha$	1/9/0.99	4/8/0.97	-	-
	1/11/0.99	-	-	-

## 結論

- 文間の単語の対応だけではなく、文の**構造情報**を考慮した最適輸送距離を提案  
**Self-Attention & Gromov-Wasserstein距離**
- 言い換え識別タスクについて、WMD との比較で性能の向上がみられた

## 展望

- Self-Attention と文の**構造情報**についての関係の明確化
- BERT の各 layer, 各 head が捉えている特徴についての考察  
 layer, head で性能が大きく異なるため
- Self-Attention と似た性質を持つ行列を用いた計算の軽量化  
**計算が大変！**

- Kusner et al. From word embeddings to document distances. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 957–966, Lille, France, 07–09 Jul 2015. PMLR. URL <http://proceedings.mlr.press/v37/kusnerb15.html>.
- Sho Yokoi et al. Word rotator’s distance: Decomposing vectors gives better representations. *CoRR*, abs/2004.15003, 2020. URL <https://arxiv.org/abs/2004.15003>.
- Ashish Vaswani et al. Attention is all you need, 2017. URL <https://arxiv.org/abs/1706.03762>.
- Jacob Devlin et al. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.
- Facundo Mémoli. Gromov-wasserstein distances and the metric approach to object matching, 2011.
- Titouan Vayer et al. Fused gromov-wasserstein distance for structured objects: theoretical foundations and mathematical properties, 2018. URL <https://arxiv.org/abs/1811.02834>.
- Titouan Vayer et al. Optimal transport for structured data with application on graphs, 2019.
- Yuan Zhang et al. Paws: Paraphrase adversaries from word scrambling, 2019.