3BPM1-06

Self-Attention行列を用いた 最適輸送距離に基づく言い換え識別

2022年09月07日

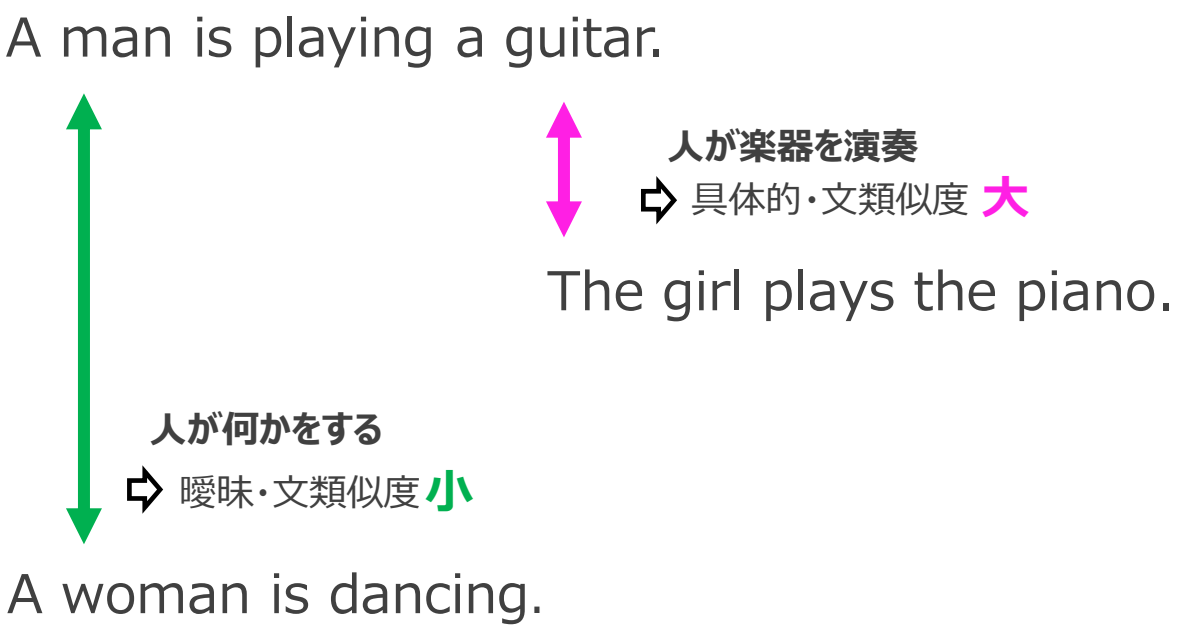
山際　宏明[1], 横井祥[2,3], 下平英寿[1,3]

[1]京都大, [2]東北大, [3]理化学研究所

目次

文類似度

文類似度 ＝ 2つの文がどれくらい似ているかを数値化



文類似度の測定

2つの文間の「距離のようなもの」を用いて**大小**を測定したい…

⇨ **最適輸送**を用いた手法が知られている

最適輸送

分布の違いを数値化

1. 分布の設定

$\alpha := \sum_{i=1}^{n} a_i \, \delta_{\boldsymbol{x}_i}$; $\{\boldsymbol{x}_i\}_{i=1}^{n}$上に重み $\boldsymbol{a} = (a_i)_{i=1}^{n}$を持つ分布, $\beta := \sum_{j=1}^{m} b_j \, \delta_{\boldsymbol{y}_j}$; $\{\boldsymbol{y}_j\}_{j=1}^{m}$上に重み $\boldsymbol{b} = (b_j)_{j=1}^{m}$を持つ分布

重み $\boldsymbol{a}$ を重み $\boldsymbol{b}$ に**輸送**したい. 輸送量 P_{ij} は? (図1)

2. コスト関数の導入

x_i, y_j 間のコスト関数 $c(x_i, y_j) =: C_{ij}$ (≥ 0)

$\boldsymbol{C} = (C_{ij})$に関する $\boldsymbol{a}, \boldsymbol{b}$ の**最適輸送**: コストの総和が最小となる割り当て方

$L_{\boldsymbol{C}}(\boldsymbol{a}, \boldsymbol{b}) = \min_{\boldsymbol{P} \in \boldsymbol{U}(\boldsymbol{a}, \boldsymbol{b})} \sum_{i=1}^{n} \sum_{j=1}^{m} C_{ij} P_{ij}$: **最適輸送距離**

$\boldsymbol{U}(\boldsymbol{a}, \boldsymbol{b}) = \{\boldsymbol{P} = (P_{ij}) \mid \forall i, j \ \sum_{j=1}^{m} P_{ij} = a_i, \sum_{i=1}^{n} P_{ij} = b_j\}$: 実現可能な割り当て

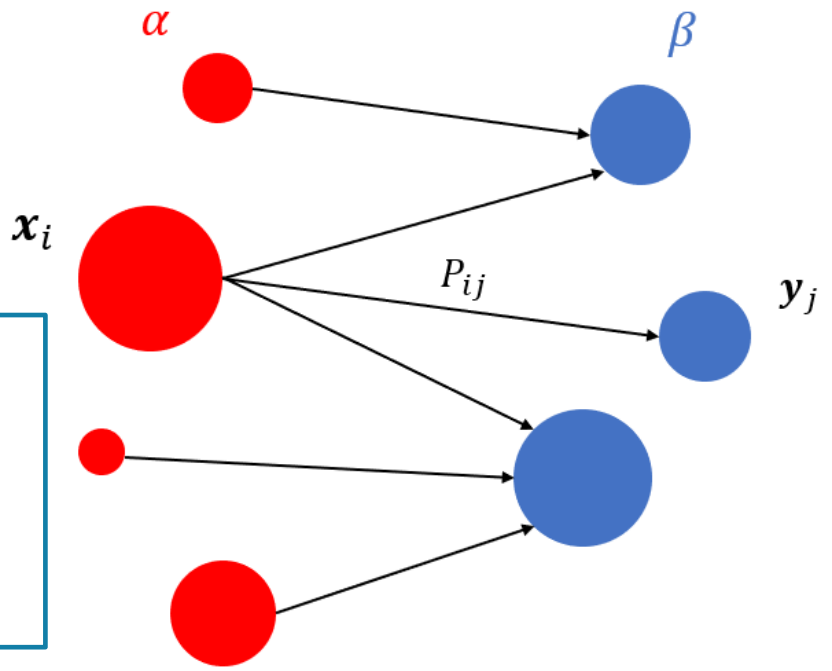


図1: 輸送量の割り当て方の例

$\boldsymbol{D}$:距離行列 として, $\boldsymbol{C} = \boldsymbol{D}^{\boldsymbol{p}} = \left(D_{ij}^{p}\right)$ の時の $L_{\boldsymbol{D}^{\boldsymbol{p}}}(\boldsymbol{a}, \boldsymbol{b})^{\frac{1}{p}}$: p-**Wasserstein 距離**

既存の最適輸送に基づく手法

最適輸送距離を用いた文類似度計算

1. 文を単語の分散表現(単語の意味を反映した数百次元の特徴ベクトル)を用いて表す: $s = [w_1, \ldots, w_i, \ldots, w_n]^{\mathrm{T}}$
2. 分散表現から文を表す分布, 単語間のコストを設定.
3. 最適輸送距離で2つの文の違いを数値化.

各単語の分散表現

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

We have much rain $(n = 4)$ It rains hard $(m = 3)$

分布:

$$\mu_s = \sum_{i=1}^{n} \frac{1}{n} \delta_{w_i}$$

コスト関数:

$$c(w_i, w_j) := \|w_i - w_j\|$$

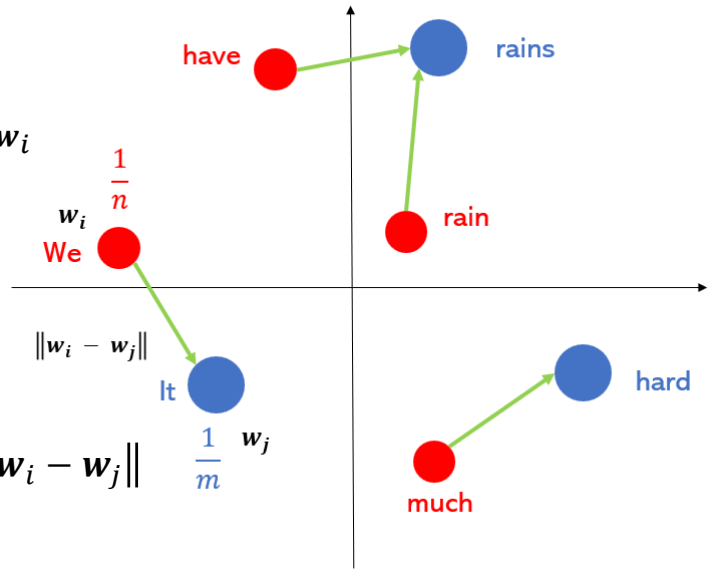


図2: WMDを用いた文類似度計算

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

We have much rain $(n = 4)$ It rains hard $(m = 3)$

分布:

$$\mu_s = \sum_{i=1}^{n} \frac{\|w_i\|}{Z} \delta_{w_i}$$

$$Z = \sum_{i=1}^{n} \|w_i\|$$

コスト関数:

$$c(w_i, w_j) := 1 - \cos(w_i, w_j)$$

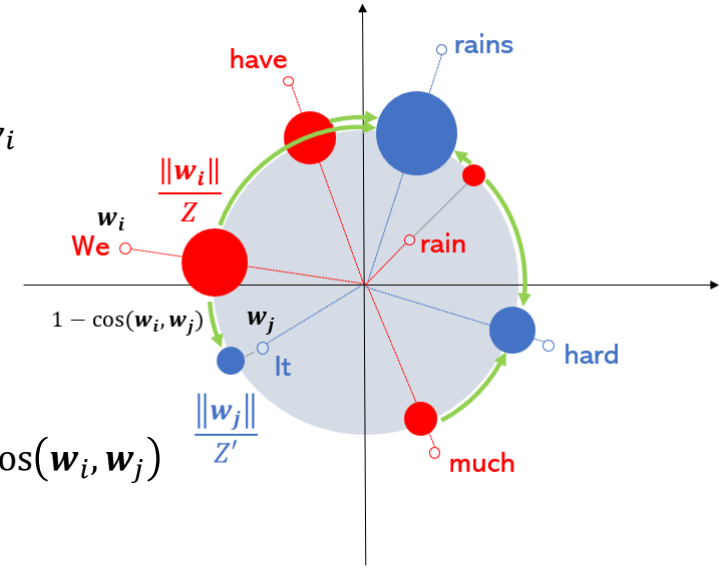


図3: WRDを用いた文類似度計算

既存手法の問題点

WMD, WRDでは文を単語の集合とみなす

⇨ 文中の**語順**を考慮しない.
(※考慮せずに上手くいくことも多い)

(例) **異なる**2つの文で WMD = 0 (WRDでも同様)

- Flights from New York to Florida
- Flights from Florida to New York

要点

- 2つの文は言い換え表現ではないが, WMD = 0
- この例では「WMD を用いた文類似度 ≠ 人間が考える文類似度」

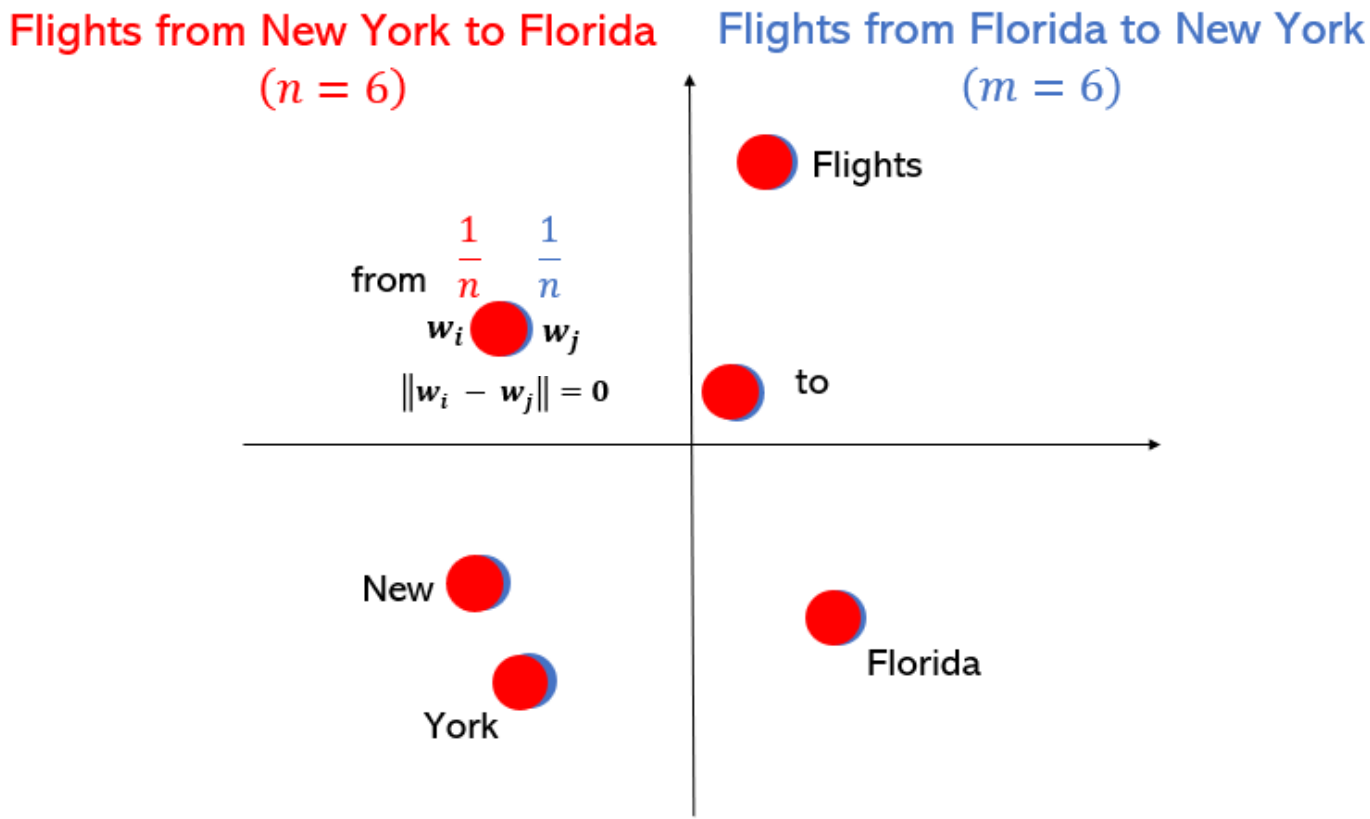


図4: 例の2文で動的な分散表現で WMDを計算する様子.

※ 動的な分散表現の場合, 分布は一致はしない.

提案手法の概要

概要

「文間の単語の対応 ＋ 各文内の単語による構造情報」となる最適輸送距離の提案

既存手法　　Self-Attention **行列** &
Gromov-Wasserstein **距離**

道具1: Self-Attention 行列

BERT [Devlin et al., 2019] の Self-Attention 行列は**構造情報**に関連 [Clark et al., 2019]

文をBERTに入力し,
取り出した Self-Attention行列 を利用

道具2: Gromov-Wasserstein 距離

Gromov–Wasserstein 距離は**構造**について焦点を当てた最適輸送距離

Self-Attention 行列

特徴

- 文を入力し，BERT などのモデルから取り出せる．
- 成分が単語同士の関係の度合いを表すスコア．

⇨ 文の構造情報に関連した行列
[Clark et al., 2019]

(例) a woman is dancing.

低 (woman, a) = 0.01

高 (woman, dancing) = 0.8

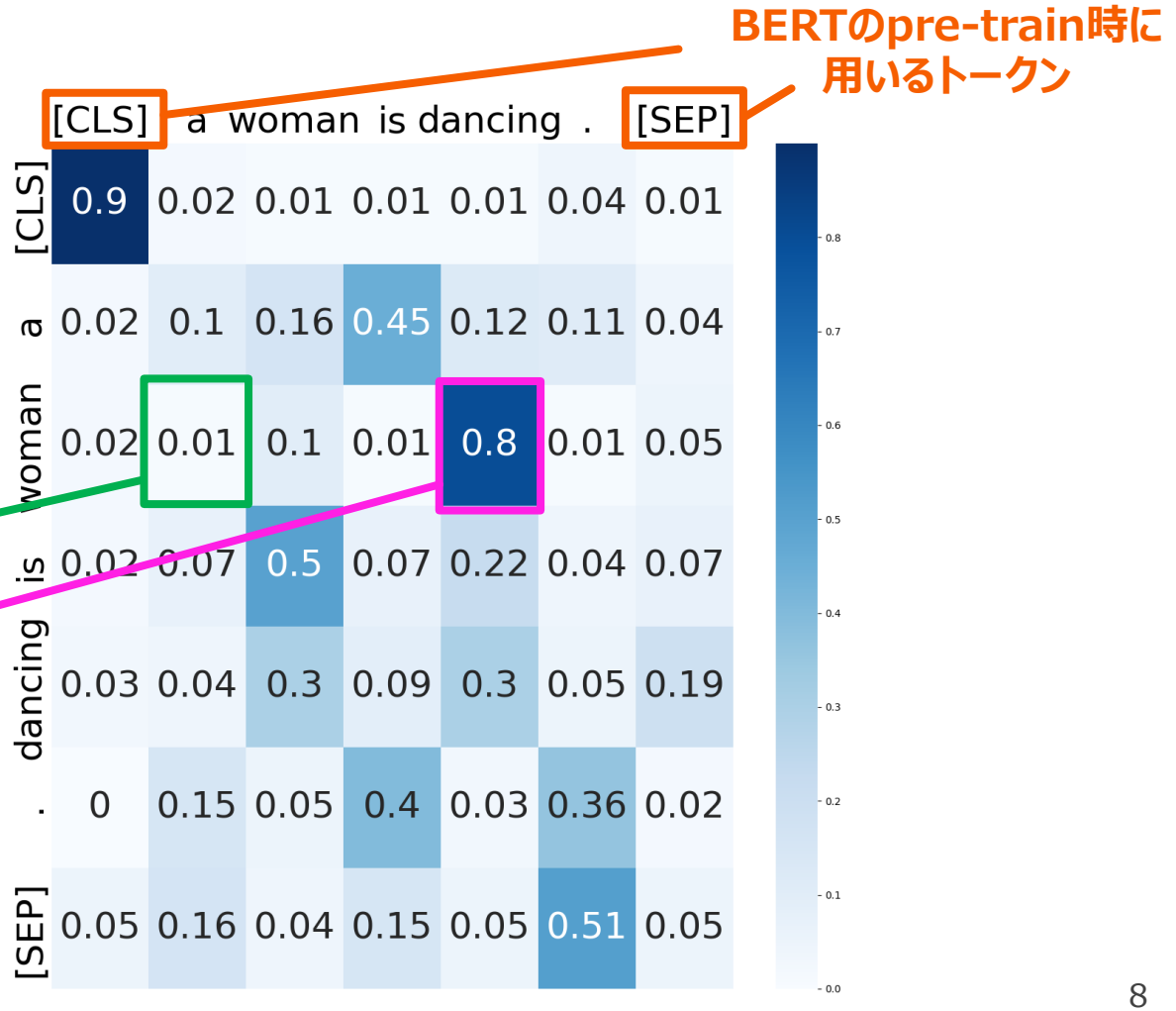


図5: “a woman is dancing.” を事前学習済みのBERTに入力した際のSelf-Attention 行列

Gromov-Wasserstein 距離

n個の要素とm個の要素についての構造行列: $\boldsymbol{C}^x = \left(C^x_{ii'}\right) \in \mathbb{R}^{n\times n}, \boldsymbol{C}^y = \left(C^y_{jj'}\right) \in \mathbb{R}^{m\times m}$

p-Gromov-Wasserstein 距離 [Mémoli, 2007]

$$p\text{-GW}(\boldsymbol{a},\boldsymbol{b}) \coloneqq \min_{P\in U(\boldsymbol{a},\boldsymbol{b})} \left(\sum_{i,j,i',j'} \left|C^x_{ii'} - C^y_{jj'}\right|^p P_{ij}P_{i'j'}\right)^{\frac{1}{p}}$$

コスト: $i \mapsto j,\ \ i' \mapsto j'$ と対応させた時, $C^x_{ii'}, C^y_{jj'}$がどれだけ変化するか

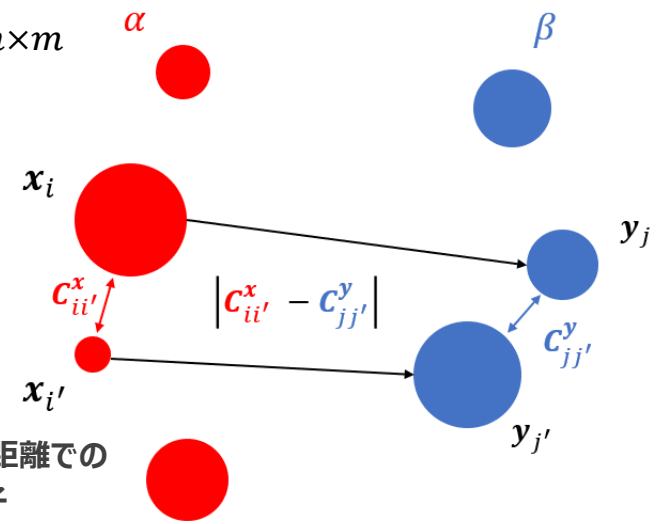


図6:
Gromov-Wasserstein 距離での
$C^x_{ii'}$, $C^y_{jj'}$の比較の様子

p, q, λ-Fused Gromov-Wasserstein 距離 [Vayer et al., 2019]

$$p,q,\lambda\text{-FGW}(\boldsymbol{a},\boldsymbol{b}) \coloneqq \min_{P\in U(\boldsymbol{a},\boldsymbol{b})} \left(\sum_{i,j,i',j'} \left\{(1-\lambda)D^q_{ij} + \lambda\left|C^x_{ii'} - C^y_{jj'}\right|^q\right\} P_{ij}P_{i'j'}\right)^{\frac{1}{p}}$$

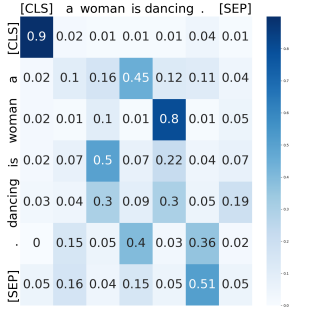
Wasserstein距離　　Gromov-Wasserstein距離

提案手法

手順

1. BERT の分散表現を用いて, WMD (WRD) から文 s, s′ の確率ベクトル $\boldsymbol{a}, \mathbf{b}$ とコスト行列 $\mathbf{C}$ を計算.
2. 事前学習済みのBERTに文s, s′を入力, Self-Attention 行列 $\mathrm{SA}^x, \mathrm{SA}^y$ を取り出す. ([CLS], [SEP] は除く)

a woman is dancing. → **input** → **pre-trained BERT** → **extract** → **Self-Attention行列**



3. $\mathrm{SA}^x, \mathrm{SA}^y$ を用いて 2 文間の最適輸送距離$d_\lambda(s, s')$を計算.

$$d_\lambda(s, s') := \min_{\boldsymbol{P} \in \boldsymbol{U}(\boldsymbol{a}, \boldsymbol{b})} \sum_{i,j,i',j'} \left\{ (1-\lambda) \frac{C_{ij}}{\sum_{k,l} C_{kl}/nm} + \lambda \left(\frac{\left| \mathrm{SA}^x_{ii'} - \mathrm{SA}^x_{jj'} \right|}{\sum_{k,l,k',l'} \left| \mathrm{SA}^x_{kk'} - \mathrm{SA}^x_{ll'} \right| / n^2 m^2} \right)^2 \right\} P_{ij} P_{i'j'}$$

- オーダーを揃えるため, 平均で規格化.
- Gromov-Wasserstein 距離の項の 2 乗は計算量を $O(n^2 m^2) \rightarrow O(n^2 m + n m^2)$ に落とすため [Vayer et al., 2019].

実験設定

提案手法の性能を**言い換え識別**タスクで性能評価

PAWS [Zhang et al., 2019]

- 重複する単語を含む2つの文は, 言い換え表現か?
- AUCで性能評価 (最大1, 最小0)

(例) 1. Flights from New York to Florida
2. Flights to Florida from NYC
3. Flights from Florida to New York

言い換え

言い換えではない

データ数. PAWS_{QQP} では, train → dev, dev → test とする.

	Train	Dev	Test	Paraphrase
$\text{PAWS}_{\text{Wiki}}$	49,401	8,000	8,000	44.2%
PAWS_{QQP}	11,988	677	–	31.3%

実験設定

- BERT は transformers [Wolf et al., 2020] の bert-base-uncased (12層12head).
- BERTの0層目 (静的), BERT の12層目 (動的, 最終層) を使用.
- $\mathrm{PAWS}_{\mathrm{wiki}}$, $\mathrm{PAWS}_{\mathrm{QQP}}$ で AUC を比較 .

baseline

bag-of-words (BOWS), 単語の分散表現の平均による文の分散表現

比較手法

- WMD, WRD, Gromov-Wasserstein 距離
- 最適輸送で語順を考慮した手法である Ordered WMD [Liu et al., 2018], WMDo [Chow et al., 2019]

実験結果

	embedding					$PAWS_{Wiki}$ AUC
baseline						
BOWS						0.492
word embeddings mean	BERT0					0.4892
	BERT12					0.5669
Ordered WMD (max iteration = 5)		metric	l_1	l_2	σ	
	BERT0	euclid	1	0.04	10	0.5934
	BERT0	cosine	12	0.2	10	**0.7304**
	BERT12	euclid	1	0.03	10	0.6573
	BERT12	cosine	3	0.075	10	0.6985
WMDo		δ				
	BERT0	19.5				0.7214
	BERT12	20				0.7212
WMD-base		layer	head	λ		
WMD (cost normalized)	BERT0					0.5838
	BERT12					0.6433
WMD-GW (cost normalized)		0	8			0.623
WMD-proposed	BERT0	0	8	0.6		0.7148
	BERT12	2	9	0.43		0.7262
WRD-base		layer	head	λ		
WRD (cost normalized)	BERT0					0.547
	BERT12					0.6118
WRD-GW (cost normalized)	BERT0	6	9			0.5264
	BERT12	2	9			0.5205
WRD-proposed	BERT0	0	8	0.75		0.7281
	BERT12	2	9	0.48		0.7156

	embedding					$PAWS_{QQP}$ AUC
baseline						
BOWS						0.4901
word embeddings mean	BERT0					0.5026
	BERT12					0.6163
Ordered WMD (max iteration = 5)		metric	l_1	l_2	σ	
	BERT0	euclid	3	0.01	10	0.583
	BERT0	cosine	10	0.06	10	0.6446
	BERT12	euclid	2.5	0.01	10	0.6517
	BERT12	cosine	1.9	0.01	10	0.6856
WMDo		δ				
	BERT0	19.5				0.5346
	BERT12	17.5				0.5337
WMD-base		layer	head	λ		
WMD (cost normalized)	BERT0					0.521
	BERT12					0.634
WMD-GW (cost normalized)		6	1			0.602
WMD-proposed	BERT0	9	1	0.4		0.7025
	BERT12	9	1	0.36		0.6972
WRD-base		layer	head	λ		
WRD (cost normalized)	BERT0					0.5078
	BERT12					0.6335
WRD-GW (cost normalized)	BERT0	10	4			0.589
	BERT12	9	11			0.5896
WRD-proposed	BERT0	9	1	0.4		**0.7121**
	BERT12	9	1	0.35		0.7019

: best : proposed : WMD (WRD), GW, proposed の比較

まとめ

- 既存手法:各要素の特徴に焦点
- 提案手法:「既存手法 + Self-Attention 行列 + Gromov-Wasserstein距離」で語順を考慮
- 提案手法の性能: PAWS データセットで向上を確認

参考文献 I

Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. 2015. From word embeddings to document distances. In Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37, ICML'15, page 957–966. JMLR.org.

Sho Yokoi, Ryo Takahashi, Reina Akama, Jun Suzuki, and Kentaro Inui. 2020. Word rotator's distance. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 2944–2960, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Facundo Mémoli. 2007. On the use of Gromov- Hausdorff Distances for Shape Comparison. In Eurographics Symposium on Point-Based Graphics. The Eurographics Association.

参考文献Ⅱ

Titouan Vayer, Nicolas Courty, Romain Tavenard, Laetitia Chapel, and Rémi Flamary. 2019. Optimal transport for structured data with application on graphs. In Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA, volume 97 of Proceedings of Machine Learning Research, pages 6275– 6284. PMLR.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase adversaries from word scrambling. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 1298–1308, Minneapolis, Minnesota. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 38–45, Online, October 2020. Association for Computational Linguistics.

Bang Liu, Ting Zhang, Fred X. Han, Di Niu, Kunfeng Lai, and Yu Xu. 2018. Matching natural language sentences with hierarchical sentence factorization. In Proceedings of the 2018 World Wide Web Conference , WWW ’18, page 1237–1246, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Julian Chow, Lucia Specia, and Pranava Madhyastha. 2019. WMDO: Fluency-based word mover’s distance for machine translation evaluation. In Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1) , pages 494–500, Florence, Italy. Association for Computational Linguistics.