P4-04

Self-AttentionとFused Gromov–Wasserstein距離に基づく文類似度計算

山際宏明[1], 横井祥[2, 3], 下平英寿[1, 3]
1. 京都大学大学院 2.東北大学大学院 3.理化学研究所革新知能総合研究センター

概要

- Wasserstein距離を用いて文の類似度(STS)をうまく計算できる. (例) Word Mover's Distance (WMD) [Kusner+'2015], Word Rotator's Distance (WRD) [Yokoi+'2020].
- 文を単語集合と見なしていて単語間の関係性（たとえば係り受け）が考慮されていない.
- Self-Attention に単語同士の関係性の情報がうめこまていることにヒントを得て, Gromov–Wasserstein距離を用いてSelf-Attention 行列同士がどれだけ似ているかを測る.

背景

文の類似度を最適輸送を用いて測ることができる

WMDでは, 単語埋め込み集合同士をマッチングするための“輸送”コストを文間の非類似度とみなした

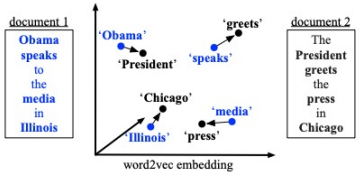

図1[Kusner+'2015]

Wasserstein距離

2つの集合の要素同士を対応づけて距離を測る尺度(図1)

$W_p(u,v)^p := \min_{T\in\prod(u,v)} \sum_{i,j} d_{ij}{}^p T_{ij}$. ここで$\prod(u,v) := \{T \in \mathbb{R}_{\geq 0}^{n\times m} : \sum_j T_{ij} = u_i, \sum_i T_{ij} = v_j\}$は$u$, vによって定まるカップリング.

WMD (文類似度計算にWasserstein距離を用いた手法)

u, vは一様分布, $d(w_i, w_j') = \| w_i - w_j' \|_2$, $WMD(s,s') = \min_{T\in\prod(u,v)} \sum_{i,j} \| w_i - w_j' \|_2 T_{ij}$

提案手法

概要

- 文間の単語の対応だけでなく各文内の単語たちがなす構造情報を考慮した距離を用いたい.
- Self-Attentionは構造情報 (各文内の 単語-単語 の結びつきの強さ) とみなせる [Ma+'2020].
- Self-Attention行列がなす構造が一致しているかをGromov–Wasserstein距離を用いて2つの文の構造がどれくらい似ているかを測ることができる.

詳細

- **Self-Attention**

TransformerのEncoderに文を入力して得られる行列.

- **Gromov –Wasserstein距離**

2つの文を表す行列同士が同じ“形”をしているか測る尺度(図2).

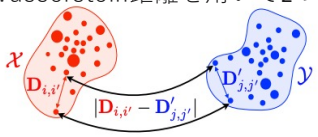

図2 [Peyré+'2020]

Xから Yの対応を決めた時, Xでの点同士の距離がYではどれくらい変化するかを見ている.
Xで近い点が, Yでも近い点になっている時, X, Yは“似た”構造をしていると考えることができる.

$GW_p(u,v)^p := \min_{T\in\prod(u,v)} \sum_{i,j,i',j'} |D_{ii'} - D_{jj'}|^p T_{ij} T_{i'j'}$

- **Fused Gromov –Wasserstein距離**

Wasserstein距離とGromov –Wasserstein距離の組み合わせ. Self-Attention行列 A, A' を用いた手法を提案. $FGW(s,s') = \min_{T\in\prod(u,v)} \sum_{i,j,i',j'} ((1-\alpha)\, d(w_i, w_j') + \alpha |A_{ii'} - A'_{jj'}|) T_{ij} T_{i'j'}$

W GW

実験

目的

「Wasserstein距離にGromov–Wasserstein距離を組み合わせることで文類似度計算の性能が向上するか？」を試してみる.

実験設定

- STS Benchmark [Cer+'2017] を使用. 人間によるgoldスコアとの相関係数で性能評価する.
- 分散表現・Self-Attentionはhuggingface/transformersのbert-base-uncasedから取り出す.
- WMDと性能を比較.
- $FGW(s,s')$ の Wasserstein距離にはWMDを用いる.

結果

表1: W(WMD), GW(手法1, 2), FGW(手法1, 2)

手法1: Self-Attention(SA)全体を平均してひとつの SA を作り、これを用いて最適輸送コストを計算・評価
手法2: 各 SA でそれぞれ最適輸送コストを計算して、評価結果全体を平均

W(WMD)	GW(手法1)	GW(手法2)	FGW (手法1)	FGW (手法2)
42.78	10.97	14.50	42.93	**43.48**

結論

WMDにSelf-Attentionを用いて構造情報を足す事で文類似度計算における精度向上がみられた.

dev セットを用いたパラメータチューニング

使用したbert-base-uncasedは12層12head.

- devセットに対し αを0.01から0.99 まで 0.01ずつ動かした時, 最大値をとるα は$\alpha = 0.99$.
- layerごとにheadについて相関係数を平均した時最大となるのはlayer id = 0.
- devセットでlayer id = 0 のうちWMDより性能が高いhead id = 0, 1, 4, 5, 6, 7, 8, 9, 11を選択.