Self-Attention 行列を用いた最適輸送距離に基づく言い換え識別

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

文類似度は, 機械翻訳などの生成モデルの損失関数や, 類似文章検索等のシーンで頻繁に用いられる重要な道具である. 自然言語には, 共通する単語をほとんど持たないが文の意味が類似する性質 (例えば, *Obama speaks to the media in Illinois.* と *The President greets the press in Chicago.*) や, 共通する単語を多く持つが文の意味が異なる性質 (例えば, *Flights from New York to Florida.* と *Flights from Florida to New York.*) があるため, 文類似度の計算は一筋縄にはいかない. そのため, 自然言語処理の分野では上手く文類似度を測るための様々な手法が提案されている. 特に最適輸送を用いて文類似度を測る手法としては, Wasserstein 距離に基づく手法 [2] が知られているが, [2] では文を単語の分散表現の集合とみなし文類似度を計算するため, 文中の語順を考慮することができない. そこで, 文内の単語間の関係を反映した最適輸送距離を考えるために (1) 事前学習された BERT[1] に文を入力して得られる Self-Attention 行列, (2) 構造の類似度を測る最適輸送距離である Gromov–Wasserstein 距離に着目した.

(1) について, 入力文に対し, BERT の各層のパラメータを用いて Self-Attention 行列と呼ばれる文内の単語間の関係を表した行列を計算することができる. $\mathbf{X} \in \mathbb{R}^{n\times d}$ を文の行列表示とし, $\mathbf{W}^K, \mathbf{W}^Q \in \mathbb{R}^{d\times d}$ を BERT のパラメータとすると, 以下で定義される.

$$\mathrm{SA}^{\mathbf{X}} := \mathrm{softmax}\Big(\frac{\mathbf{X}\mathbf{W}^Q(\mathbf{X}\mathbf{W}^K)^\top}{\sqrt{d}}\Big) \in \mathbb{R}^{n\times n} \tag{1}$$

(2) について, Wasserstein 距離は, 比較する分布の要素間での距離を用いた比較を行い, 各要素の特徴に焦点を当てる. Gromov–Wasserstein 距離は, 同じ分布内の要素間での構造情報を用いた比較を行い, 構造に焦点を当てる. また Wasserstein 距離と Gromov–Wasserstein 距離の両方を考慮した Fused Gromov–Wasserstein 距離 [3] は, 要素の特徴と構造の両方に焦点を当てた割り当てとなる. 文の組について, 行列表示を $\mathbf{X} \in \mathbb{R}^{n\times d}, \mathbf{Y} \in \mathbb{R}^{m\times d}$, 確率ベクトルを $\mathbf{a} \in \mathbb{R}^n, \mathbf{b} \in \mathbb{R}^m$, 文間の距離行列を $(C_{ij}) \in \mathbb{R}^{n\times m}$, 文内の構造行列を $(C^{\mathbf{X}}_{ii'}) \in \mathbb{R}^{n\times n}, (C^{\mathbf{Y}}_{jj'}) \in \mathbb{R}^{m\times m}$ とし, $\mathbf{U}(\mathbf{a},\mathbf{b})$ を割り当ての集合とすると, $p, q \geq 1, \lambda \in [0,1]$ を用いて

$$\mathrm{FGW}_{p,q,\lambda}(\mathbf{a},\mathbf{b}) := \min_{\mathbf{P}\in\mathbf{U}(\mathbf{a},\mathbf{b})} \Big\{ \sum_{i,j,i',j'} \big((1-\lambda)C_{ij}^q + \lambda|C^{\mathbf{X}}_{ii'} - C^{\mathbf{Y}}_{jj'}|^q\big)\, P_{ij}\, P_{i'j'} \Big\}^{1/p} \tag{2}$$

$\lambda = 0$ のとき, Wasserstein 距離, $\lambda = 1$ のとき Gromov–Wasserstein 距離に対応する.

本研究では, 既存手法で用いられる文間の単語の対応に加え, 事前学習後の BERT から得られる入力文の Self-Attention 行列を Gromov–Wasserstein 距離に構造行列として用いることで, 語順によって意味が変わる場合でも文類似度を上手く計算する手法を提案する. 具体的には (2) を用いて次の最適輸送距離を計算する.

$$\widetilde{\mathrm{FGW}}_\lambda(\mathbf{a},\mathbf{b}) := \min_{\mathbf{P}\in\mathbf{U}(\mathbf{a},\mathbf{b})} \sum_{i,j,i',j'} \left\{ (1-\lambda)\frac{C_{ij}}{\sum_{k,l} C_{kl}/nm} + \lambda\left(\frac{|\mathrm{SA}^{\mathbf{X}}_{ii'} - \mathrm{SA}^{\mathbf{Y}}_{jj'}|}{\sum_{k,l,k',l'} |\mathrm{SA}^{\mathbf{X}}_{kk'} - \mathrm{SA}^{\mathbf{Y}}_{ll'}|/n^2m^2}\right)^2 \right\} P_{ij}\, P_{i'j'} \tag{3}$$

参考文献

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, volume 1. ACL, 2019.

[2] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37. PMLR, 2015.

[3] Titouan Vayer, Laetitia Chapel, Rémi Flamary, Romain Tavenard, and Nicolas Courty. Fused gromov-wasserstein distance for structured objects: theoretical foundations and mathematical properties. *CoRR*, abs/1811.02834, 2018.