文間意味的類似度のベンチマークタスクと実応用タスクの乖離

阿部香央莉[*1] 横井祥[*1*2] 梶原智之[*3] 乾健太郎[*1*2]　1. 東北大学 2. 理化学研究所 3. 愛媛大学

概要
- 意味的類似度ベンチマーク STS ⇔ 実応用タスクの間で**評価の乖離**が生じている
  - 原因：STS側の文長の短さ、語彙の簡単さ など

→ STSが実応用タスクに向けた意味的類似度ベンチマークとして効果を発揮していない
- 意味的類似度ベンチマークのあり方の見直しが必要

背景：STS◎ → NLP実応用◎？

- 2文間の意味的類似度予測は**多くのNLP実応用で必要**

[Severym+'13, Lan+'18, Liu+'19]

例1: 機械翻訳評価

(hyp, ref) → Model → *(hyp, ref)*の類似度スコア

例2: 関連文検索

(query, passage) → Model → *(query, passage)*の関連度スコア

例3: 自動採点

(student response, ref) → Model → *(student's res, ref)*の類似度スコア

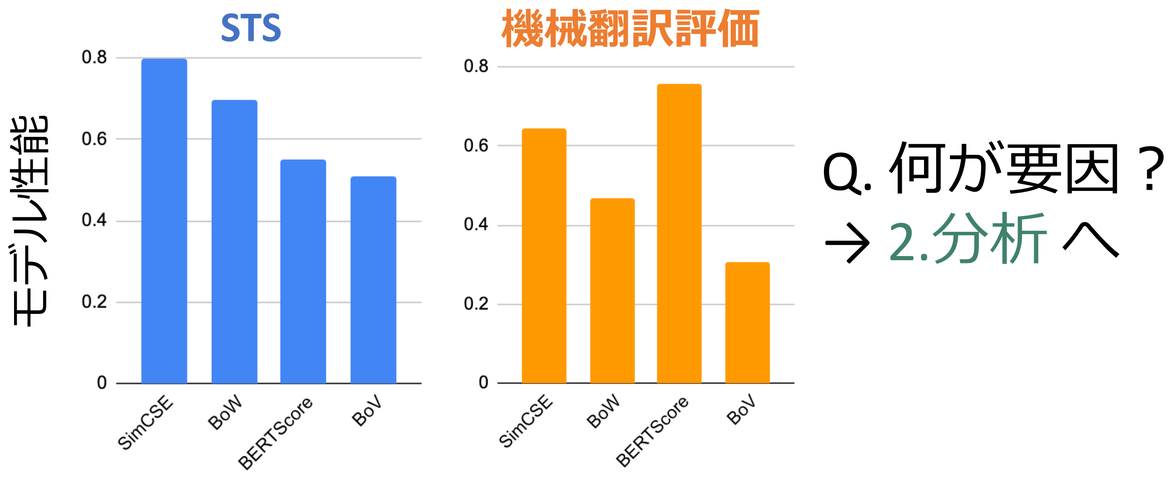
- 「意味的類似度ベンチマーク**STSが解ける → NLP実応用の性能向上**」という**仮定**が**分野のコンセンサス**
  - STSで評価されたモデル [Conneau+'17, Logeswaran&Lee'18, Cer+'18] が**過去に機械翻訳評価で活躍** [Shimaoka+'18]
  - STSモデルを**言語生成モデル学習時に用いて性能向上** [Wieting+'19][Yasui+'19]
  - STSでの評価における**競争的なモデル提案** [Reimers&Gurevych'19, Zhang+'20, Giorgi+'21, Gao+'21]

しかし，この仮定は正しいのか？

本研究は以下を分析：
1. **STS⇔実応用タスク間の評価の乖離の実証**
2. **評価の乖離の要因を分析**

1. 検証：STS ⇔ 実応用タスク間の評価乖離

- **STS**⇔**実応用タスク**で各意味的類似度予測モデルの**性能、順位が変動 = 評価の乖離**あり
  → 「STSが解ける → 実応用性能向上」とは限らない？



Q. 何が要因？
→ 2.分析 へ

実験設定

データセット (3種類)

意味的類似度ベンチマークタスク
- **STS-b** [Cer+'17]

NLP実応用タスク
- **機械翻訳評価: WMT17** [Bojar+'17]
- **関連文検索: MS-MARCO** [Bajaj+'18]

モデル (15種類)

- BoW (2)
- BoV (6)
- BERTScore (6) [Zhang+'20a]
- SimCSE (1) [Gao+'21]

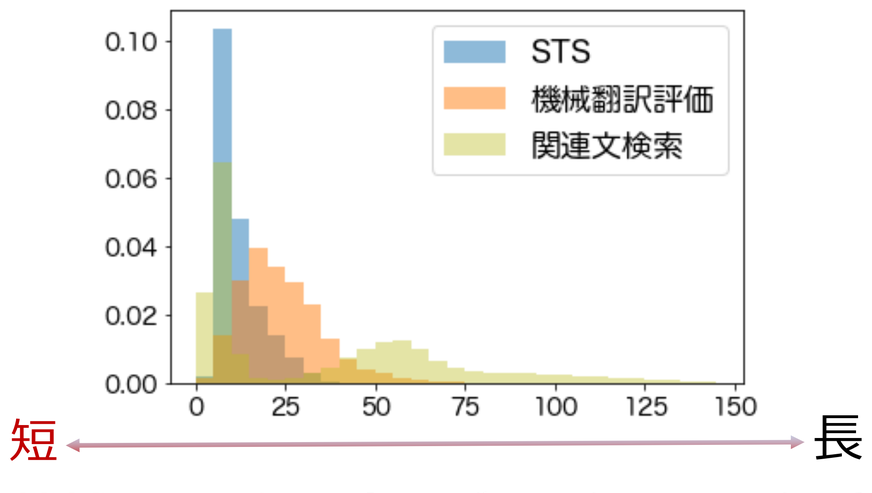
※事前学習モデル (Vec, LM) やプーリングのvariants

各モデルの**STSでの性能** ⇔ **実応用タスクでの性能のスピアマン相関**を計測 → **値が低い = 評価の乖離大**

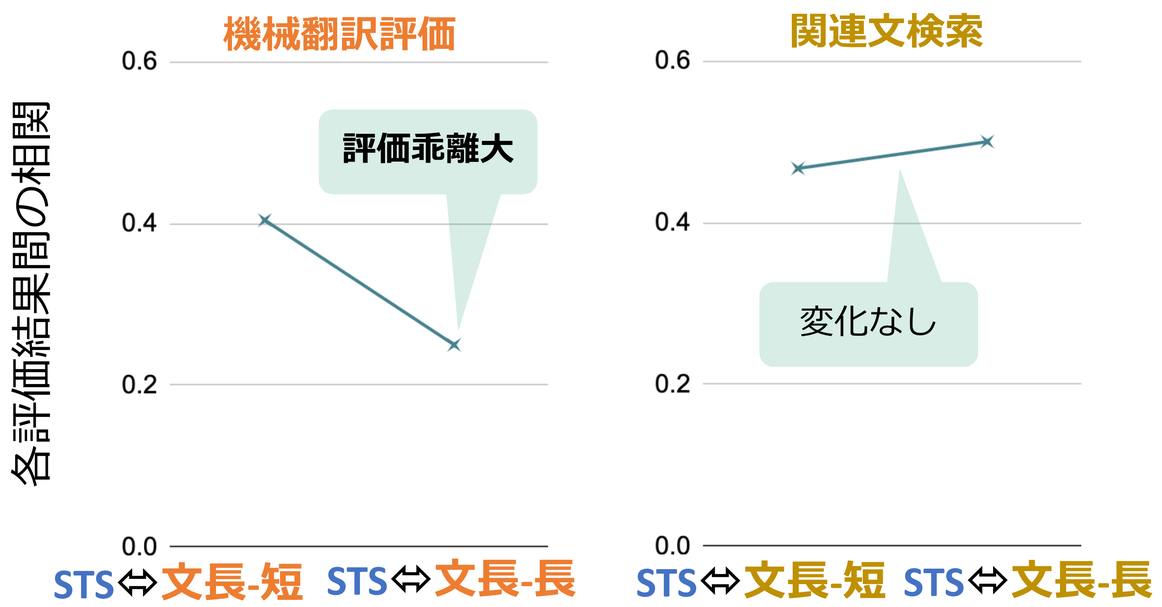
2. 分析　仮説① STSの文長が短すぎる → 評価乖離？

- **STS**⇔**実応用タスク**間には文長分布の差あり
  - STSは文長が非常に**短い**

各タスクデータの文長（文中の単語数）のヒストグラム



- **STS**⇔**機械翻訳評価**の**文長が長い例との評価乖離大**
  = **STSの文長が短すぎる**のが要因？



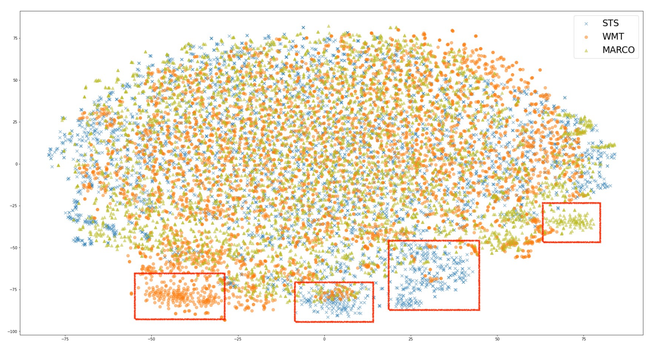
仮説② STSの語彙分布がずれている → 評価乖離？

- **STS**⇔**実応用タスク**の間には語彙分布の差あり
  - STSは**簡単**な語彙で構成

	STS	機械翻訳評価	関連文検索
avg. 単語長*	**6.97±2.76**	7.34±2.83	10.1±4.83

* 単語長が短いほど低難易度とみなす [Kincaid+, 1975]

- STS ⇔ 実応用タスクの**一部単語表現クラスタのずれ**



- **STS**⇔**両応用タスクのSTS語彙割合が低い例との評価乖離大**
  = **語彙分布が離れている**のが要因？