

楽天レビューデータの商品属性を利用した文の分散表現の学習

NII-IDRユーザーフォーラム2020
スタートアップセッション S04
谷 健太郎, 上野 史, 太田 学 (岡山大学)

背景・目的

文の分散表現は分類/ラベル付け/特徴抽出等に利用でき、**機械学習モデル (Sentence-BERT等)**で生成可能。



楽天データセット

レビューとその評価, 対象の商品属性 (商品名, 商品ジャンル, 店舗等)がある。



【楽天商品レビューの利用法】

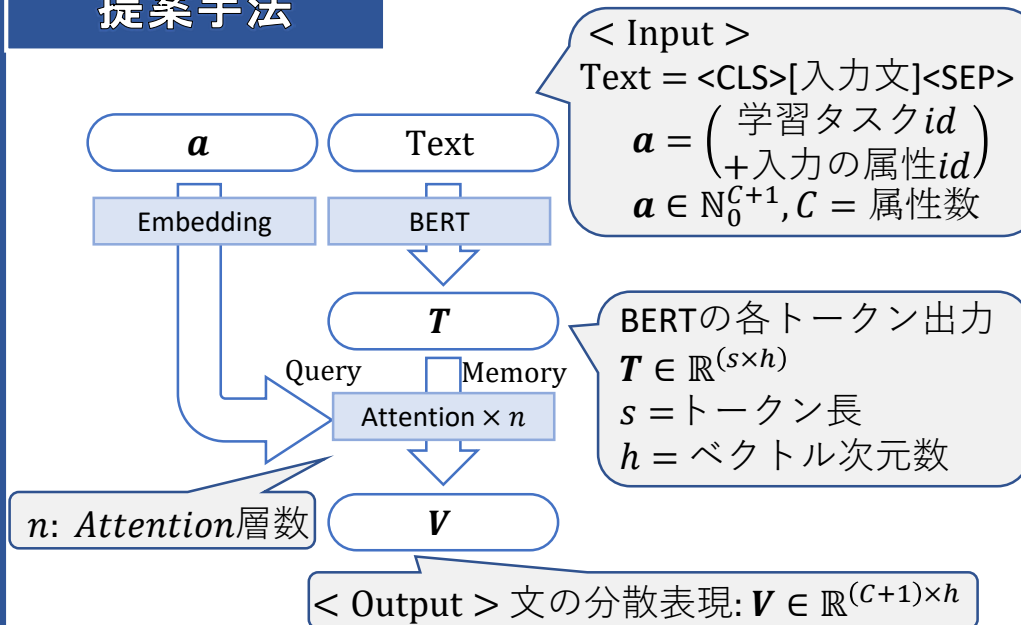
入力文: レビュー内容

属性: 投稿者ID, 評価ポイント, 店舗名, 商品名, 商品ジャンル

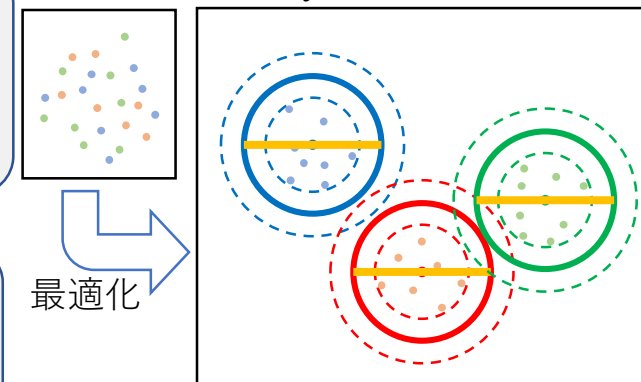
入力文の複数属性を利用したモデルと学習法の提案



提案手法

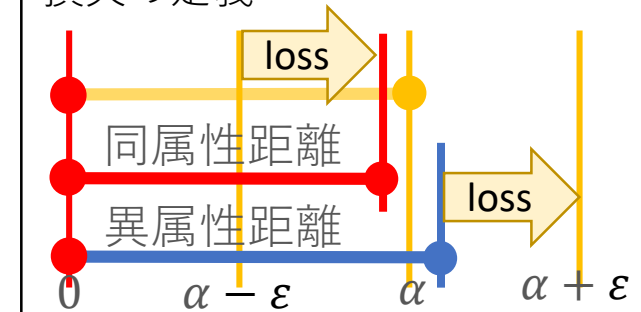


分散表現 V_i の分布の最適化



同色の点: 属性 a_i が同じ文の分散表現

損失の定義



【楽天商品レビューの類文判別実験】

データセット: 「商品情報, みんなのレビュー・口コミ情報」

- 分類[A]: 正例
- 分類[B/C/.../F]: 負例
分類の組み合わせ検証
- 学習・検証・評価用に分割
- 交差検証
- ジャンル等の異なる組み合わせでの実験

ジャンル*	商品名	評価	分類
一致	一致	一致	A
		不一致	B
	不一致	一致	C
		不一致	D
不一致	/	一致	E
		不一致	F

*商品ジャンル

今後の課題

1. モデル, 損失関数等の実装
2. 楽天商品レビューでの類文判別の実験
3. ジャンル, 評価ポイント推定等の実験