進捗報告

1　今週やったこと

- BERT-Transformer モデルで多値分類をした.
- 14 クラス分類では精度が向上しなかったため、それぞれのクラスにおいて, そのクラスに属するか, それ以外の 13 クラスに属するかの 2 値分類をした.
- BERT-Transformer モデルに, 各ラベルと関連度の高いフレーズを抽出するターゲットフレーズラベリング器の実装中

2　データセット

2.1　楽天トラベルレビュー : アスペクトセンチメントタグ付きコーパス

楽天トラベルレビュー : アスペクトセンチメントタグ付きコーパスとは, 楽天グループ株式会社が提供しているデータセットである. 日本語レビュー文章とそれぞれの文章について, 立地, 部屋, 食事等の 7 項目のアスペクトに対するポジティブまたはネガティブのタグが付与されている. 総データ数は 76624 で, 朝食, 夕食, 風呂, サービス, 施設, 立地, のポジティブ, ネガティブの 14 クラスである. 今回は 14 のいずれのラベルにも属さないデータを除くことで, 少なくとも 1 つのラベルに属し, 語彙数が 10 以下と 100 以上のデータを取り除いたデータ群にした. 総データ数は 50211 である. 表 1 にデータの具体例を示す.

3　実験

3.1　各クラス 2 値分類

楽天トラベルレビュー : アスペクトセンチメントタグ付きコーパスを用いて多値分類をした. 訓練データ数を 5600, バリデーションデータを 1400 として学習をした. 14 クラスの分類ではなく, 14 クラスそれぞれのクラスにおいて, そのクラスかそれ以外かの各クラス 2 値分類をした. 表 2 に学習時のパラメータを示す. 多値分類の時と異なるのは Transformer 層の出力数や損失関数である.

表 2: 各クラス 2 値分類のパラメータ

パラメータ	値
BERT 層の入力次元数	768
BERT 層の出力次元数	768
Transformer 層の層数	1
Transformer 層の入力次元数	768
Transformer 層の出力次元数	2
バッチサイズ	20
最適化関数	Adam
学習率	0.0001
損失関数	CrossEntropyLoss
エポック数	15

只今実験が滞っているため, 結果は報告できない. 現時点でわかっていることは, 以下の通りである.

- 1 エポック毎の学習量が多く, 使用できるデータ数が限られる.
- 訓練時の検証データでの予測正解率がうまく推移せず詰まってしまっている. (全てのクラスで正解率が変動しない.)
- 何度か学習率を変更したり, プログラムを書き換えたりしてみたが, どうしてもうまくいかなかった.

表 1: 両方のラベルが立っているデータの具体例

テキスト	朝食 po	朝食 ne	夕食 po	夕食 ne	風呂 po	風呂 ne
お部屋も広くて、お料理もとても美味しく、部屋の露天風呂からは 星がプラネタリウムのように広がっていて、とにかく最高でした。	1	0	1	0	1	0
部屋も綺麗で、対応もよく、朝食もおいしいので とても満足しています。	1	0	0	0	0	0
気になるところは廊下の天井が低いのと、部屋数がたくさんあり、 温泉が集中するとお風呂まちになるところくらいですかね。	0	0	0	0	0	1