進捗報告

1　今週やったこと

- 実験に用いるデータセットの調整をした.
- いくつかの調整をして, ポジネガ両方のラベルが立っているデータと語彙数が 15 以下と 100 以上のデータを取り除いた.

2　データの内容

調整を行ったデータセットは少なくとも 1 つのラベルが立っているデータ群である. 表 1 に朝食のポジネガクラスを読み込んだデータにおいて, 同じクラスでポジティブとネガティブのラベルが両方立っているデータ数や, 平均, 語彙数などの値について示す. 語彙数は東北大学の BERT を用いて, データの日本語文章から tokenizer で獲得した.

次にデータセットの調整内容について示す.

- 1. 全文が英語で書かれたレビュー文を取り除いた.
- 2. そして, 日本語文字を全角にし, 英単語を小文字に統一した.
- 3. また, 数値は今回のタスクにおいて有用ではないと判断し, 全てを 0 に変換し, 各種記号については取り除いた.

以上の手順を踏まえた後にポジネガ両方のラベルが立っているデータと語彙数が 15 以下と 100 以上のデータを取り除いた. これらの処理を終えると 53192 個のデータが 43920 になった. ストップワードを取り除くことも検討したが一度今回作成したデータセットでの実験結果を見てからにすることを考えた.

3　実験

3.1　実験 1

まず初めにデータ内容の 1　3 の項目について調整を行ったデータを用いた実験 1 をした. 訓練データ数を 3120, バリデーションデータを 1020 として学習をした. データに含まれるラベルが 0 が多く, 評価指標として正解率を用いると高くなってしまい正しく評価できないと考えたので F1 値を用いた. 表 2 に学習時のパラメータを示す.

表 2: 多値分類タスクに用いたパラメータ

値	パラメータ
入力層の次元数	768
出力層の次元数	14
バッチサイズ	14
最適化関数	Adam
学習率	1e-4
損失関数	BCEWithLogitsLoss

続いて図 1 に実験 1 の損失を示す. また, 図 2 に実験 1 における学習時のバリデーションデータの F1 値の推移を示す.

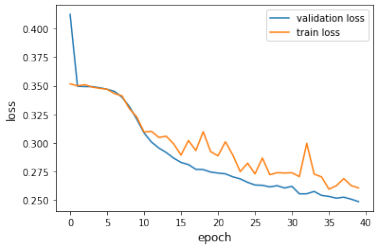

図 1: 実験 1 の損失の推移

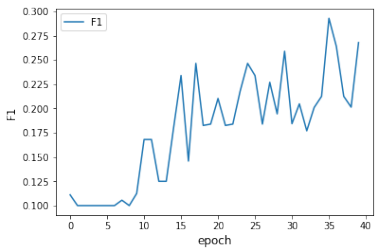

図 2: 実験 1 の F1 値の推移

図 1 からは学習が進んでいることが確認できるが図 2 からは F1 値があまり上昇せず, 安定していないことがわかる.

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	サービス po	サービス ne	立地 po	立地 ne	設備 po	設備 ne	部屋 po	部屋 ne
立地、最上階、部屋からの景色、エアウィーヴ等の良い点と比較しても、次の機会に泊まるかは疑問です.	0	0	0	0	0	0	0	1	1	0	1	1	1	1
外観を見て失敗したと思いましたが、中に入ると別世界でした.	0	0	0	0	0	0	0	0	0	0	1	1	0	0
古いながらも大変メンテナンスされていますので清潔でした.	0	0	0	0	0	0	0	0	0	0	1	1	1	1
食事も夕・朝とも質量ともに問題なかったのですが,逆に朝は量が多すぎるくらいでした.	1	1	1	0	0	0	0	0	0	0	0	0	0	0

3.2 実験 2

同様のパラメータで, データ内容の 3 の処理を行わないデータを用いて実験 2 をした. 訓練データ数を 3120, バリデーションデータを 1020 として学習をした. 実験 2 でも同様の理由で F1 値を評価指標として用いた. 図 3 に実験 2 の損失を示す. また, 図 4 に実験 2 における学習時のバリデーションデータの F1 値の推移を示す.

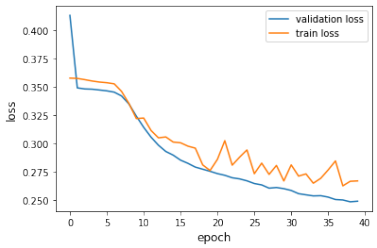

図 3: 実験 2 の損失の推移

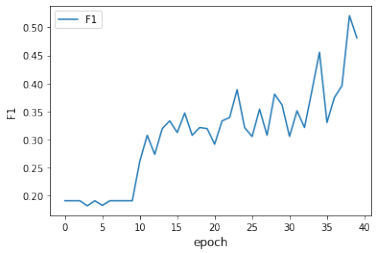

図 4: 実験 2 の F1 値の推移

図 3 からは学習が進んでいることは確認できるが少し過学習に陥っている可能性があるように見受けられる. また図 4 からは図 2 と比較してバリデーションデータでの F1 値は高くなっていることがわかる.

実験 1, 2 を比較すると用いたデータにおける処理が結果に大きな影響を与えたことがわかった. 特にデータの内容の 3 は数値や各種記号が実験に不要な情報だと考えた故にした処理だったが, その仮定が誤っていたと考えられる. 次回以降の実験では, よりデータの調整は慎重にするべきであると考えた.