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.

- 全文が英語で書かれたレビュー文を取り除いた.
- 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.
- 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.

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　実験

43920 のデータからランダムにバッチサイズ (14) の倍数分取り出して学習を行い, テストデータを用いた予測を行おうとしている.