進捗報告

1 今週やったこと

- Sentence-BERT の勉強

2 Sentence-BERT の勉強

前回の実験でわかっていなかったところ，足りなかったところを調べた．追加の実験はできていない．

2.1 データセット

京都大学の JSNLI データセット[*1]を用いて学習した．JSNLI データセットは自然言語推論の標準ベンチマークである SNLI(Standord Natural Language Inference) を日本語に翻訳したものであり，ラベル，前提，仮説の 3 つ組を表している．前提及び仮説は JUMAN++ によって形態素分割されている．ラベルは entailment，contradiction，neutral の 3 種類である．表 1 にこのデータセットの基本情報を示す．表 2 にラベル毎のデータ数を示す．

表 1　JSNLI データセットの基本情報

訓練データ数	テストデータ数	平均文長	最大文長
533005	3916	13.1	94

表 2　JSNLI データセットのラベルごとのデータ数

ラベル	entailment	contradiction	neutral
訓練データ	176309	178700	177996
テストデータ	1432	1156	1328

表 3 に表 1 のデータを元に作成した訓練データ，テストデータ，類似文章ペアの数を示す．訓練データに関しては参考となるコードを読んでみたものの，いまいちよくわからなかった．処理としては前提，仮説の組からラベルを参考に損失関数に合う形の anchor，positive，negative の組のデータを作成しているはずである．テストデータはデータの前提，仮説から一致しない文章のみを集めたデータである，類似文章ペアはラベルが entailment のもののみを抽出したデータでありテストデータとともに評価関数に入力するデータである．

表 3　作成したデータセット

訓練データ	テストデータ	類似文章ペア
294579	5809	1432

[*1] https://nlp.ist.i.kyoto-u.ac.jp/index.php?%E6%97%A5%E6%9C%AC%E8%AA%9ESNLI%28JSNLI%29%E3%83%87%E3%83%BC%E3%82%BF%E3%82%BB%E3%83%83%E3%83%88

2.2 損失関数と評価関数

損失関数は MultipleNegativesRankingLoss を使用した．入力の形式は anchor，positive，negetive の 3 つのデータの組となっており，各 anchor に対し，対応する positive との距離が小さく，それ以外の positive および negative との距離が大きくなるようにする損失関数である．バッチサイズ分のデータを考慮して計算される．

評価関数には ParaphraseMiningEvaluator を使用した．ParaphraseMiningEvaluator は，まず文書集合に含まれるすべての文章ペアで類似度を計算する．そしてスコアが高いペアから順に見ていき，与えられた類似文章ペアに含まれていれば正解と評価して Average Precision と 最良の F1 スコア及びその閾値を求める評価関数である．入力は文書集合と類似文章ペアである．今回はラベルが entailment のものを類似文章ペアとした．

2.3 実験

表 3 に実験に用いたパラメータを示す．表 4 に実験結果を示す．ウォームアップ後と比較して Average Pricision と F1 値は上昇しており，学習はできていると考えられる．

表 4　学習に用いたパラメータ

parameter	value
train	294579
test	5809
learning rate	2×10^{-5}
optimizer	AdamW($\beta_1 = 0.9$, $\beta_2 = 0.999$)
loss	MultipleNegativesRankingLoss
evaluator	ParaphraseMiningEvaluator
epoch	1
batch size	48

表 5

モデル	Average Precision	Pricision	Recall	F1
ウォームアップ後	9.33	15.09	27.93	19.60
訓練後	13.04	18.88	33.31	24.10

3 次にやること

- 訓練データの処理についてもう少し調べる
- 京大の BERT で同様に動かして，tokenizer の違いによる影響を調べる