進捗報告

1 今週やったこと

- 実験
- 漢字の成分に関する調査

2 データセット

中華灯謎データベースから答えが一文字の灯謎 79727 問を収集した.

そして元々の問題文-正解ペアに対し, 同じ問題文-不正解のペアを作成した. 故に実験用データは 79727 問から 159454 問に拡張した.

これらのデータセットに正解は 1 , 不正解は 0 のようにラベルを付け, 灯謎の答えが正解か不正解かを判明するように実験した.

対照実験するため, データセットを正解, 不正解 1 対 2 の比率でデータを拡張した.

2.1 データオーグメンテーション

データセットの形について, 正解は問題と正解の字でペアで, 不正解は同じ問題と違う字のペアである.

具体的には図 1 のように正解の答えを 1000 位下にずれで, 問題と不正解のペアで作ります.

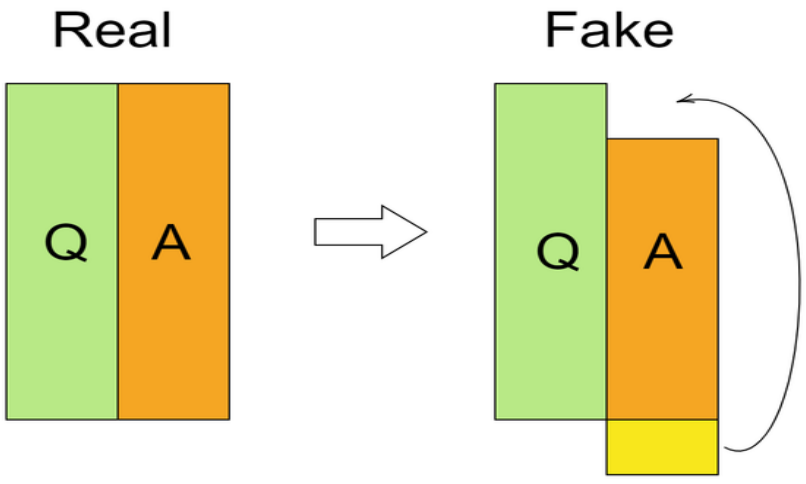

図 1: データオーグメンテーション

データの数は表 1 のように示す.

表 1: Dateset

experiment	Real	Fake	Total	Training	Validation	Testing
LSTM(baseline)	79727	79727	159454	127563	15946	15945
LSTM(Augmentation)	79727	159454	239181	191345	23918	23918
GRU (bert base chinese)	79727	79727	159454	127563	15946	15945

3 実験

今週は hugging face に公開された bert base chinese モデル [1] で実験した.
先週の LSTM に対し, データを正解, 不正解 1 対 2 の比率で対照実験した.

3.1 実験用モデル

今回の実験は LSTM モデルを使用した.
対照実験として,LSTM (baseline),LSTM (データ数 1 対 2 拡張),GRU (bert base chinese)3 つの方法で 200 epoch の実験を行った.
そして今回のモデル構造について,3 つの実験は全部二層の双方向 RNN 構造で, Embedding 次元数を全部 300 に設定した.

3.2 実験結果

実験結果として,LSTM (baseline) に対し,Train Loss と Validation Loss は 0.363 と 0.847 に収束し,Train Accuracy と Validation Accuracy は各自 83.25 と 66.96 パーセントに収束した. LSTM (データ数 1 対 2 拡張) に対し,Train Loss と Validation Loss は 0.292 と 0.534 に収束し,Train Accuracy と Validation Accuracy は各自 87.20 と 80.28 パーセントに収束した.
一方,GRU (bert base chinese) に対し,Train Loss と Validation Loss は各自 0.469 と 0.831 に収束し,Train Accuracy と Validation Accuracy は各自 76.40 と 57.19 パーセントに収束した.
実験結果は図 2 , 図 3 , 図 4 のように示す.
Testing データによる結果は表 2 のように示す.

表 2: Testing result

experiment	TP	TN	FP	FN	Accuracy	Precision	Recall
LSTM(baseline)	5564	2386	5530	2466	49.88	50.15	69.29
LSTM(Augmentation)	4277	14484	1521	3191	80.30	75.64	59.67
GRU (bert base chinese)					55.19		

GRU (bert base chinese) 実験のプログラムは修正中で, 完成してから Testing データの結果を補足する.
結論として, データを正解, 不正解 1 対 2 の比率で設定した方が表現が良いである.

4 漢字の成分に関する調査

「文字には有益な情報が詰め込まれている !? 」[2] という論文から見ると, 日本語の文字と sub-文字に対するデータセットは GlyphWiki, IDS, KanjiVG, KRADFILE 四つが存在している.
この中に IDS は中国語の漢字も含まれ, 直接実験に使用することも可能になる.

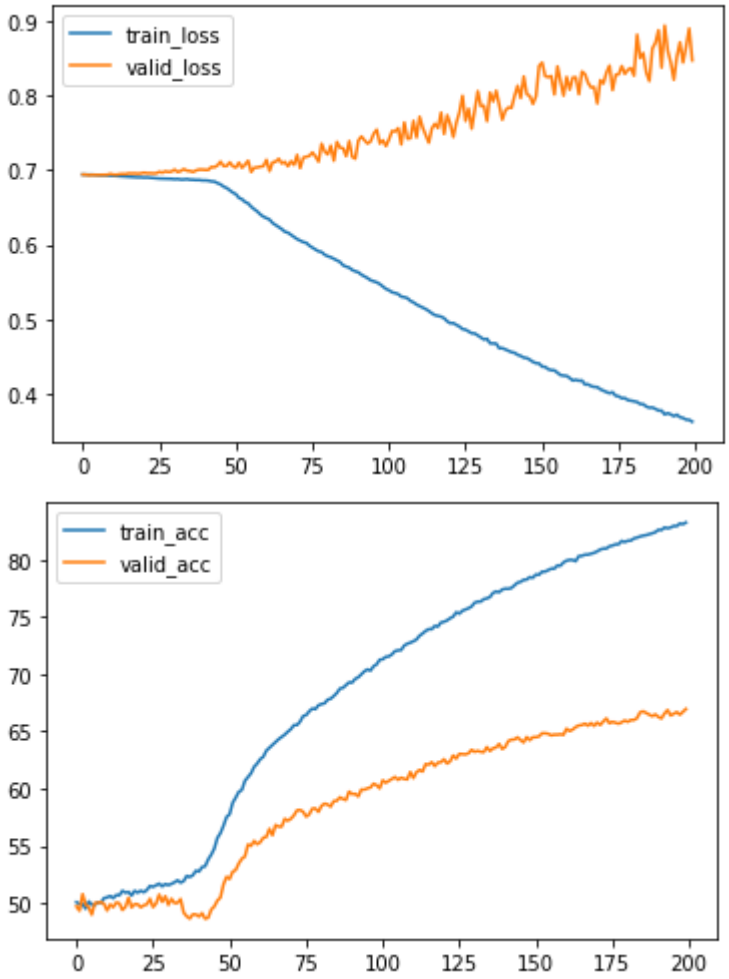

図 2: LstmBaseline

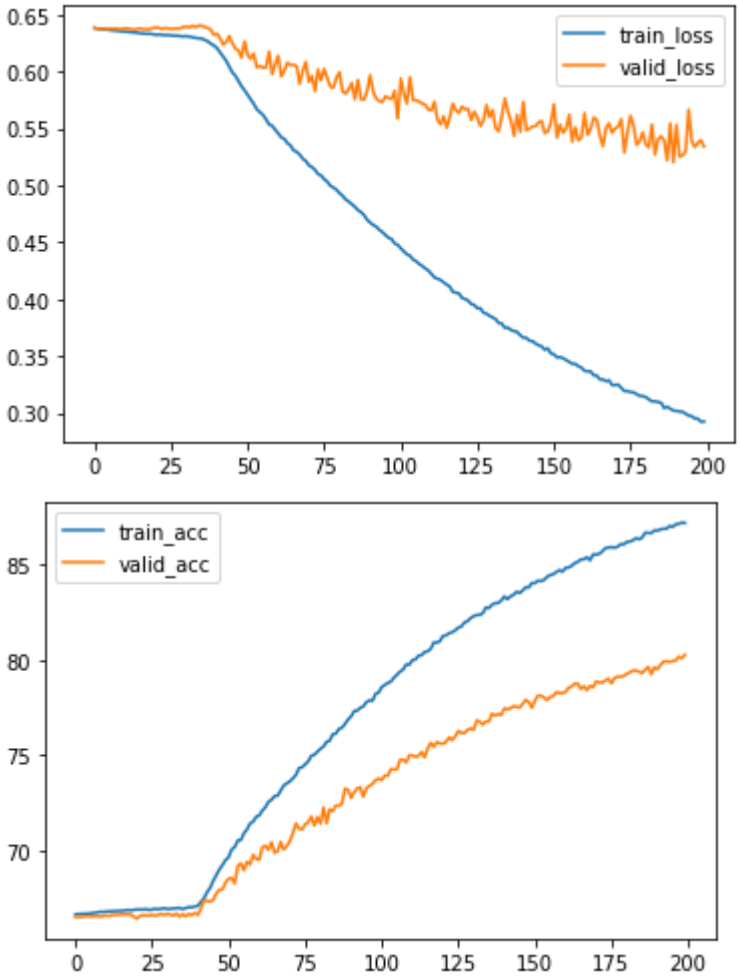

図 3: LstmDataAugmentation

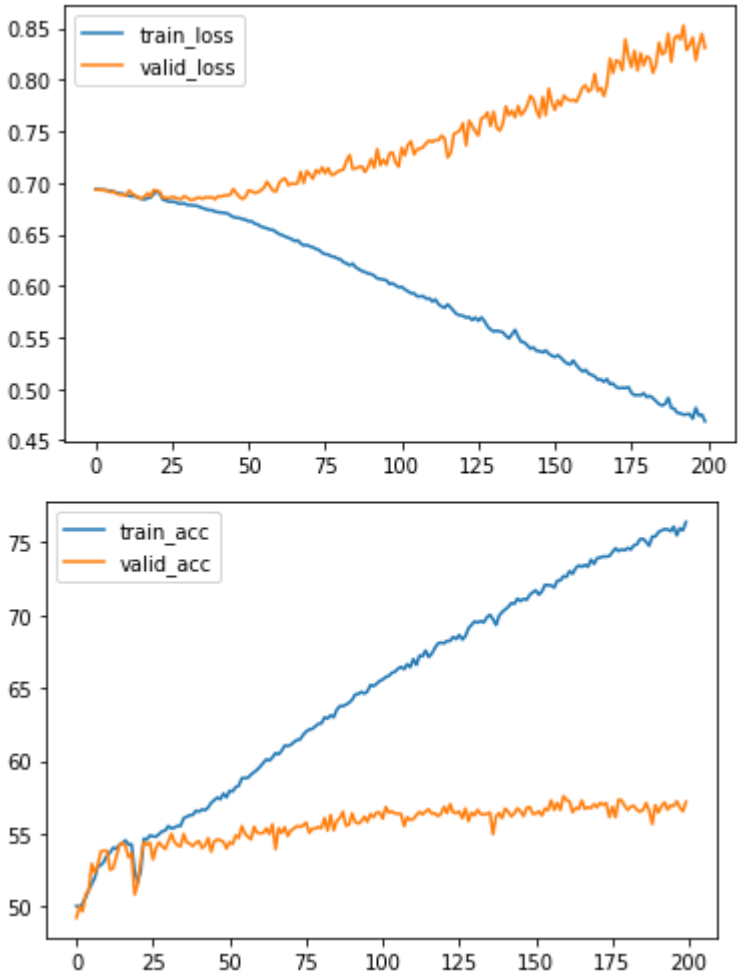

図 4: BertBaseChinese

4.1 次のステップ

次のステップについて, まず IDS を利用し, 現在の実験の精度を上回る可能性があるかを確認する.
次は答えを生成するモデルに関する取り組む. 図 5 にモデルの構造 (発想) を示す.

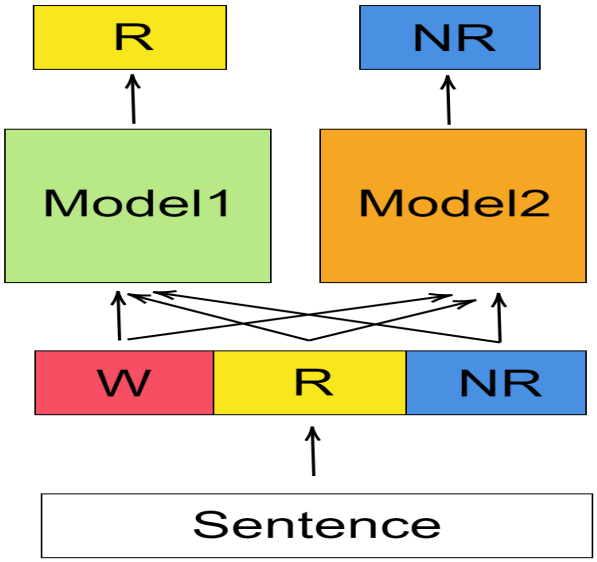

図 5: BertBaseChinese

5 来週目標

- IDS で実験すること
- 他の Bert モデル検討すること

参考文献

[1] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R 辿 miLouf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics.

[2] `https://ai-scholar.tech/articles/treatise/text-ai-181`.