進捗報告

1 今週やったこと

- データ処理
- Seq2Seq モデルでの実験

2 データセット

2.1 IDS データセットの補充

論文 [1] により, 漢字は大まかに分ける Shallow Mode と細かいに分ける Deep Mode 二つの分け方がある.

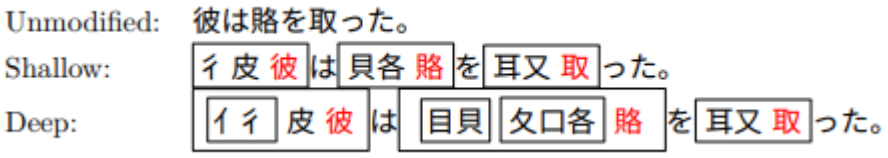

図 1: 漢字の分け方 [1]

今回実験は IDS の Shallow Mode を利用したが, ある sub 漢字は表現できない問題がある.(例えば漢字爽の x 型成分はこのまま表現できない)

これらの表現できない成分をより細かい画にわけ, 書き順で補充した.

図 2: データ補充例

2.2 灯謎データセットの処理

「算式で解決できる問題」は「答えの漢字の成分は問題の中 (漢字或いは漢字の成分) にある」と定義した.

以上の定義により, 最初に集まった灯謎 79725 問を算式問題 (True) 57535 問と非算式問題 (False) 22190 問に分け, 算式問題だけ利用した.

表 1: Testing result

Data Type	All	True	False
Number of Data	79725	57535	22190

3　実験

3.1　実験用モデル

今回は Seq2Seq モデルで実験した.

対照実験をするため, GRU で漢字から sub 漢字 (GRU_Char2Sub), GRU で sub 漢字から sub 漢字 (GRU_Sub2Sub), Attention で漢字から sub 漢字 (Attn_Char2Sub), Attention で sub 漢字から sub 漢字 (Attn_Sub2Sub), 四つの方法で実験した.

3.2　実験結果

実験結果として, GRU_Char2Sub 実験の Train Loss と Validation Loss は 0.138 と 6.346 に収束し, GRU_Sub2Sub 実験の Train Loss と Validation Loss は 0.106 と 6.821 に収束した.

そして Attn_Char2Sub 実験の Train Loss と Validation Loss は 0.454 と 6.049 に収束し, Attn_Sub2Sub 実験の Train Loss と Validation Loss は 1.090 と 5.682 に収束した.

実験結果は図 3 , 図 4 , 図 5 , 図 6 のように示す.

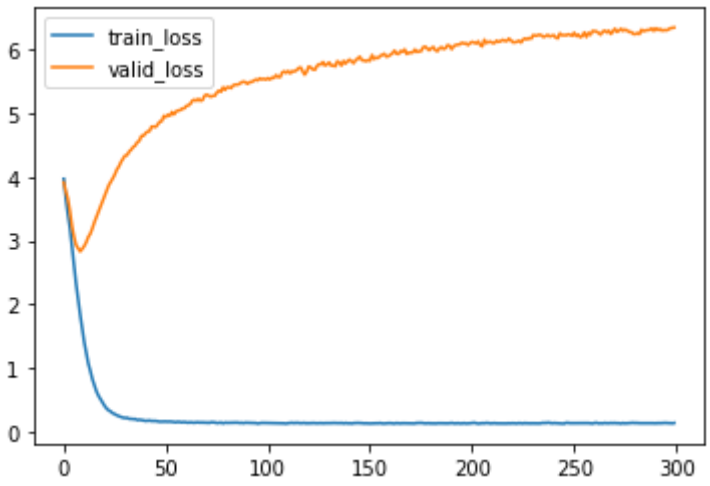

図 3: GRU_Char2Sub の Train Loss と Validation Loss

Test データ検証用コードはまだ作成中である.

4　来週目標

- Test データ用コードを完成すること

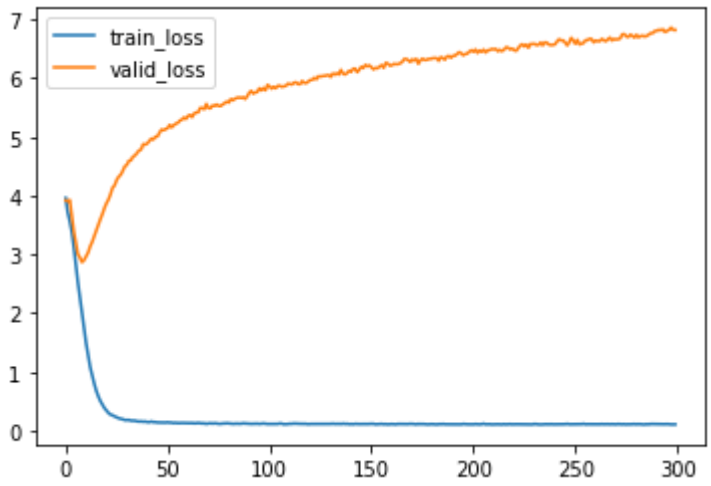

図 4: GRU_Sub2Sub の Train Loss と Validation Loss

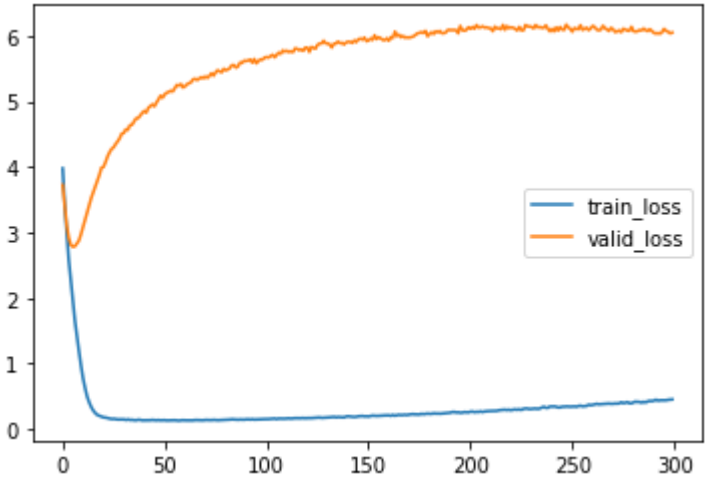

図 5: Attn_Char2Sub の Train Loss と Validation Loss

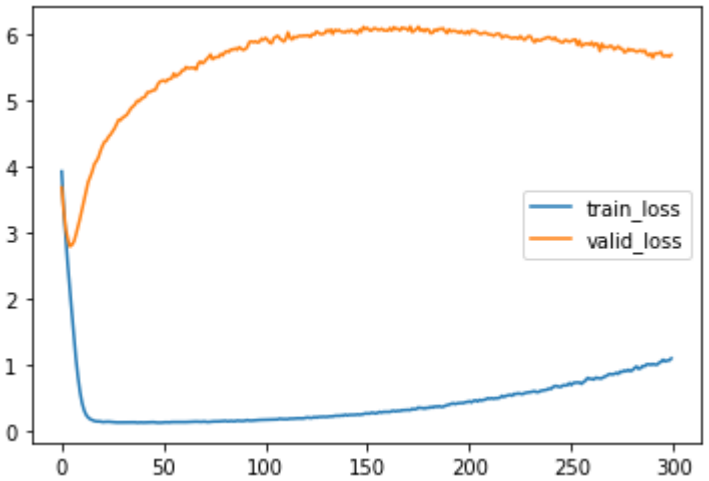

図 6: Attn_Sub2Sub の Train Loss と Validation Loss

参考文献

[1] Viet Nguyen, Julian Brooke, and Timothy Baldwin. Sub-character neural language modelling in Japanese. In *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, pp. 148–153, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.