令和 4 年 06 月 13 日 (水) ゼミ資料　　M2 陳偉斉

進捗報告

1　今週やったこと

- データセットの補足

2　データセット処理

前回であった漢字の表示されていない部分 (漢字「爽」の締める部分など) 問題について, IDS をもう一度チェックした結果として, これらの部分は実に存在している. 故に今までデータ不足の部分を補足する.

作成中の IDS データセットは「漢字」,「SUB 漢字 (Shallow)」,「SUB 漢字 (Deep)」,「画」の形である.

この中に,「SUB 漢字 (Shallow)」,「SUB 漢字 (Deep)」,「画」部分は手動で補足する必要がある.

現在「SUB 漢字 (Shallow)」,「SUB 漢字 (Deep)」部分は完成したが, 「画」部分はあと 3846 字の情報が足りないため, 補足は水曜日まで完成する予定である.

3　Accuracy 計算方法

Cao らの研究 [1] により, 漢字の画は「横」「縦」「左はらい」「点 (右はらい)」「折れ」五つの類に分類できる. 今回の実験もこの五つの「画」を利用する.

書き順を考えないた, Decoder で出力した「画」を「[' 横' = num, ' 縦' = num, ' 左はらい' = num, ' 点' = num, ' 折れ' = num]」の形で出力する.

例えば漢字「春」は「[' 横' = 5, ' 縦' = 1, ' 左はらい' = 1, ' 点' = 1, ' 折れ' = 1]」の形で表示する.

個の形で実験 Accuracy を計算する.

Accuracy 計算方法は, 予測した漢字の「画」数が本物の漢字の「画」より多い場合, (Accuracy = あってる「画」数 / 予測した漢字の「画」数) であり, 予測した漢字の「画」数が少ない場合, (Accuracy = あってる「画」数 / 本物の漢字の「画」数) と定義する.

4　実験

実験は同じモデルに対して「単語」,「単語 + 漢字」,「単語+漢字 + SUB 漢字 (Shallow)」,「単語+漢字 + SUB 漢字 (Deep)」,「単語+漢字 + SUB 漢字 (Shallow) + SUB 漢字 (Deep)」,「漢字」,「漢字 + SUB 漢字 (Shallow)」,「漢字 + SUB 漢字 (Deep)」,「漢字 + SUB 漢字 (Shallow) + SUB 漢字 (Deep)」九組の対照実験を行う予定である. GRU Seq2Seq モデルは既に完成されているため, データセットが完成次第に実験できる.

5　資料

発表会資料とパワーポイントは作成中である. 今週中完成する予定である.

参考文献

[1] Shaosheng Cao, Wei Lu, Jun Zhou, and Xiaolong Li. cw2vec: Learning chinese word embeddings with stroke n-gram information. *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32, No. 1, Apr. 2018.