進捗報告

1 今週やったこと

- DCAI 原稿作成
- Manga109, 4 コマストーリデータセットの整形
- 単語出現頻度の確認

1.1 DCAI 原稿作成

8 割くらい完成．残りは JSAI の abst 提出以降に完成させる予定.

1.2 Manga109, 4 コマストーリデータセットの整形

Manga109 のうち，4 コマ漫画を抜き出して画像とセリフが対応している json ファイルを作成．BERT(事前学習済み) を用いて得たセリフの分散表現データも同様に作成.

表 1 に用いたタイトルと各タイトルの画像枚数を示す (セリフがないものあり，かっこ内はノイズデータ削除後の件数).

表 1: タイトルと件数

タイトル	件数
総件数	2932(2802)
幼稚園ぼうえい組	368
徹さん	715(585)
OL ランチ	831
高校の人達	938
少年漫画タッチ	80

1.3 単語分布

表 2 にデータセットの単語出現分布を示す

表 2: 単語出現分布

タイトル	総単語数	平均単語数	標準偏差	最も出現頻度の高いもの 3 件
幼稚園ぼうえい組	1360	3.546	11.212	...(218), の (195), は (113)
徹さん	2275	3.678	24.721	!(1010), の (318), を (260)
OL ランチ	2730	3.506	17.309	・(556), の (379), は (225)
高校の人達	3343	4.576	23.386	!(718), の (561), に (354)
少年漫画タッチ	302	2.838	4.192	...(33), の (31), ?(29)

2　来週の予定

JSAI の abst に向けてひとまず実験を回す．終わり次第、DCAI.