4 コマ漫画ストーリー データセットの拡張に関する協力のお願い

ソフトウェアシステム研究グループ
B4 高山

要旨

・現状におけるデータセットの問題

1. データ数の少なさ
   - 1 タッチにつき, 10 話ずつ (1 タッチにつき, 約 200 個のデータ)

2. アノテーション不備 (UNK ラベル)
   - セリフデータの約 3% を占める UNK データの扱いが問題

[解決策]
・**人手による新しいセリフデータの作成 (今回お願いする内容)**
・Manga109 やその他データセットを併用した半教師あり学習

アノテーション方法

1. Google Drive 上にある, moe_augmentation_tmp.csv をダウンロードする.

2. この資料 9 ページ以降に示すフキダシ内のセリフの情報を抜いた 4 コマデータにはセリフが入るべきところに id が振られている. (id:0 ~ id:98)

3. csv ファイルのそれぞれのセリフに該当する列に各自が考えたセリフと
そのセリフ全体を表すのに相応しい感情ラベルを 1 つずつ付けていく.
感情ラベルは以下の 7 種類から最も近いものを選ぶ.

ニュートラル　喜楽　驚愕　恐怖　悲哀　憤怒　嫌悪

4. Google Drive の同階層に自分用のフォルダを作成し, アノテート完了したファイル名を moe_augmentation_***.csv (*** : フォルダ名) とし, アップロードしてください.
フォルダ名は被らなければ問題ないです.

登場人物紹介

Ａくん

Ｂさん

その他

・Ｂさんらしき後ろ姿の別人女性
・アイドル

注意

- 1 つの 4 コマ内での物語の一貫性を保つことを心がけてください.
(他の 4 コマ同士は, 別時空で考えても同時空で考えても良いです.)

- 登場人物の名前には “A くん”, “B さん” は用いず, 好きな名前を付けて使用してください.
(1 つの 4 コマ内で名前が変わったりすることがないようにだけお願いします.)

- 誰のセリフかを間違えないようにお願いします. (資料 p.6 ~ 7 参照)

- 句読点は用いず, 「、」は 1 個分のスペースで表してください.
(“...” や “!”, “?” などの常識的な記号は可)

- すべてのセリフ id についてアノテートしてもらう必要はありませんが,
4 コマ単位でアノテート抜けがないようにご協力お願いします.

- フキダシの面積とセリフの量は深く気にする必要はありません.

csv ファイル 説明

	A	B	C	D	E	F	G	H	I
1	id	story_main_num	story_sub_num	koma	who	inner	speaker	what	emotion
2	0	0	0	0	B	FALSE	B		
3	1	0	0	0	A	FALSE	A		
4	2	0	0	1	B	TRUE	B		
5	3	0	1	1	A	FALSE	A		
6	4	0	1	2	B	FALSE	B		
7	5	0	1	3	B	FALSE	B		
8	6	0	1	3	A	FALSE	A		
9	7	0	1	3	B	FALSE	B		

・id : セリフに対して固有の値です.
・story_main_num, story_sub_num : オリジナルのデータにおける話数を表しています.
無視していただいて構いません.
・koma : 何コマ目のセリフかを表します.
・who, speaker : 主体と話者を表します. 基本は一致していますが, 具体的に後述します.
・inner : 実際に発せられているセリフかどうか (端的に言えば, もくもくフキダシかどうか)
・what : **セリフの内容 (記入してもらう)**
・emotion : **セリフに対する感情ラベル (記入してもらう)**

who, speaker について

・図 1 のように, 基本的にはセリフを考えている人 (who) と喋っている人 (speaker) は一致している.

・しかし, 図 2 のように A くんの脳内イメージの B さんのセリフがある場合は, who = A に対して, speaker = B となる.

<図 1> case 1

<図 2> case 2

csv ファイル 記入例

自分で登場人物の
名前を決める

	A	B	C	D	E	F	G	H	I
1	id	sto	sto	koma	who	inner	speaker	what	emotion
2	0	0	0	0	B	FALSE	B	太郎君 週末の予定は？	喜楽
3	1	0	0	0	A	FALSE	A	研究の進捗を生みますよ	ニュートラル
4	2	0	0	1	B	TRUE	B	本当かなぁ？	ニュートラル
5	3	0	1	1	A	FALSE	A	花子さん こんにちは	ニュートラル
6	4	0	1	2	B	FALSE	B	週末は随分と楽しんだみたいだね	嫌悪
7	5	0	1	3	B	FALSE	B	太郎君 研究しないと留年するよ？	嫌悪
8	6	0	1	3	A	FALSE	A	分かってはいるのですが…	悲哀
9	7	0	1	3	B	FALSE	B	一緒に卒業しようね？	悲哀

青枠で囲まれた 2 列について
記入をお願いします.


id:17
id:16
id:18


id:20
id:19


id:21


普段は…
id:22


id:9
id:8
id:10


id:11
id:12

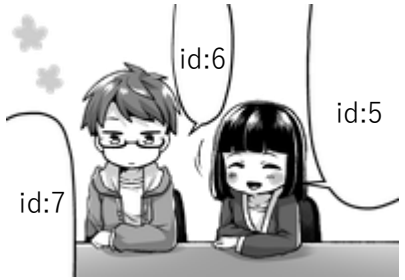

id:13
id:14


後日
ぐちゃあ
id:15


ワーーッ


id:3
ガラッ


id:4

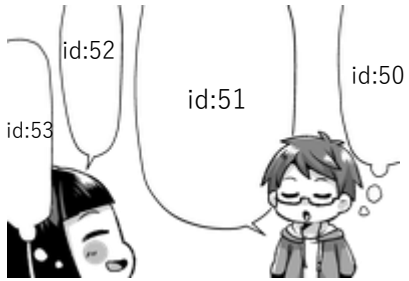

id:6
id:5
id:7

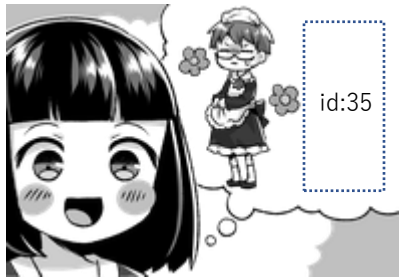

id:1
id:0


id:2


ザッ
LOV


ワーーッ





*補足 セリフの位置が分かりにくい所は [ ] で囲っています.

あと 1 ページあります


id:71
id:72


id:74
id:73
id:75


id:77
id:76


id:79
id:78


id:63
id:64


id:66
id:65


id:68
id:67


id:70
id:69


id:56
id:55


id:58
id:57


id:60
id:59


id:61
id:62


id:92
id:93
ガーン


C研究室
id:94
ハッ


id:95
ガチャ

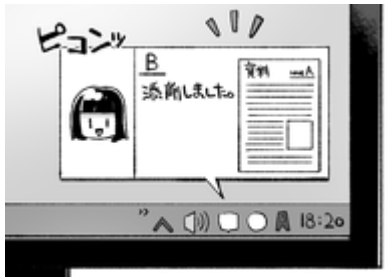

id:96
id:97
id:98


id:86
こそこそ


id:87
スタスタ


id:88
C研究室
ハッ
id:89


id:90
id:91
ガーン

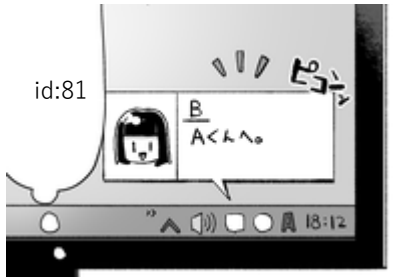

ピコンッ
B
添削しました。
18:20


id:83


id:84


id:85
ピッ


id:80


id:81
ピコン
B
Aくんへ。
18:12


id:82


ピコン
B
添削しました。
18:20