進捗報告

1 あらすじ

研究から逃げた.

2 進捗

- 未知語率再測定

3 データセットに含まれる未知語率

BERT の事前学習済モデル[1]のボキャブラリーに含まれているかどうかを各タッチについて拡張前後で算出した. また, transformers ライブラリ内のトークナイザでサブワードに分割した場合も算出した. 用いている BERT のモデルは形態素をサブワードに分割したものを基本単位として事前学習しているので, 後者の方が有用である可能性が高い.

3.0.1 形態素 (Juman++)

従来, 実験で用いていた手法における未知語率を表したのが表 1 である. 拡張前で約 25%, 拡張後で約 68% であった.

表 1: データセットに含まれる未知語率

		ギャグ	少女	少年	青年	萌え系
拡張前	総単語数	270	289	274	276	276
	未知語率	0.244	0.263	0.230	0.236	0.232
拡張後	総単語数	3030	3209	3089	3154	3154
	未知語率	0.681	0.688	0.681	0.682	0.682

3.0.2 サブワード化

3.1 未知語 (ギャグオリジナル サブワード)

```
'A', '?', 'DVD',
'B', '!', 'GPU',
',', '.'
```

[1] http://nlp.ist.i.kyoto-u.ac.jp

表 2: データセットに含まれる未知語率 (サブワード化)

		ギャグ	少女	少年	青年	萌え系
拡張前	総単語数	311	331	315	316	316
	未知語率	0.026	0.021	0.019	0.022	0.025
拡張後	総単語数	2633	2728	2664	2705	2708
	未知語率	0.090	0.089	0.088	0.089	0.089

対策として, 句読点はストップワードとして除外していたが, '.' や ',' も除外しておく. 'A' や 'B' はキャラクタ由来のワードであるから, 辞書内に含まれる固有名詞で置換する. 他のタッチのオリジナルデータでは '2' など数字も未知語として表れていた. この対策としては 数値を表すトークンまたは '0' に置換する.

3.2 未知語の品詞別頻度

萌えタッチについて, 拡張前後・サブワードにするかしないかの 4 パターンについて, 未知語を MeCab を使って品詞解析し, 品詞ごとに頻度をまとめたものが表 3 である.

表 3: 未知語の品詞別頻度 (萌えタッチ)

moe	オリジナル	拡張後	オリジナル (subword)	拡張後 (subword)
名詞	42	1927	8	258
形容詞	2	186	0	24
副詞	2	202	0	15
動詞	14	495	0	27
助動詞	14	93	0	10
助詞	9	195	0	33
連体詞	2	15	0	0
接頭詞	0	89	0	7
フィラー	1	6	0	0
接続詞	1	4	0	1
感動詞	5	7	0	0
記号	0	9	0	1

4　今後の実験予定

- サブワード化した状態で前回と同じ実験を回す.
- 直前 n - 1 文 を考慮した n 文を入力して末尾入力の感情推定をする.