進捗報告

1 あらすじ

BERT[1] pretrained モデルでの未知語数が分かると Data Augmentation 改善案が浮かぶかもしれない.

2 進捗

- 正例ラベルを変えた時の感情推定
- オリジナル ・ 拡張後における未知語数

3 正例ラベルを変えた時の感情推定

これまでの実験では正例ラベルとして喜楽としていたが, 驚愕・ニュートラルに変えて先週と同様の実験を行った. 実験の結果を表 1 に示す. 特に, 正例ラベルを驚愕とした場合は正例とまったく判断されなかった. これは, 訓練時点でラベル比が極端に正例が少なく, そもそも特徴を学習できていなかった可能性が高い.

4 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. その結果が表 2 である.

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

4.1 未知語 (ギャグタッチオリジナル)

’んだろう’, ’ございます’, ’内緒’,
’クールだ’, ’ね〜’, ’顔色’,
’大丈夫です’, ’昨晩’, ’手伝える’,
’じゃあ’, ’添削’, ’なあ’, ’あー’,
’どうぞ’, ’キャー’, ’ベタ’,
’なんで’, ’おかず’,
’はんぶん’, ’こし’, ’よっか’,
’ジャーン’, ’パフェ’, ’ました〜’,
’器用だ’, ’いろんな’, ’飽き’,
’独り占め’, ’頑張る’, ’よそ’,
’人違い’, ’恥ずかしい’,
’ただいま〜’, ’かえり’

4.2 問題点

今まで, Juman++ を用いてトークナイズすればいいと思って BERT への入力にそのまま使っていたが, もしかしたら更に BERT の トークナイザーを用いてサブワードに分けないといけなかったかもしれないので, やり直してきます. といえども, 未知語率は相当高い.

表 2: 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	0.215	0.239	0.208	0.210	0.203
拡張後	総単語数	3030	3209	3089	3154	3154
	未知語率	0.679	0.687	0.679	0.680	0.680

表 1: result

	model	ギャグ			少女漫画			少年漫画			青年漫画			萌え系			5 タッチ平均		
		Acc	Recall	F1	Acc	Recall	F1	Acc	Recall	F1	Acc	Recall	F1	Acc	Recall	F1	Acc	Recall	F1
驚愕	BERT (Last Layer)	0.758	0.000	0.000	0.806	0.000	0.000	0.859	0.000	0.000	0.662	0.647	0.500	0.719	0.000	0.000	0.761	0.129	0.100
ニュートラル	BERT (Last Layer)	0.379	0.400	0.281	0.881	0.000	0.000	0.578	0.788	0.658	0.677	0.200	0.276	0.625	0.267	0.250	0.628	0.331	0.293
喜楽	BERT (Last Layer)	0.833	0.400	0.421	0.567	0.579	0.603	0.797	0.083	0.133	0.800	0.357	0.435	0.656	0.455	0.476	0.731	0.375	0.414
ベースライン		0.85	0	0	0.43	0	0	0.81	0	0	0.78	0	0	0.66	0	0	0.71	0	0

5 Data Augmentation 改善案

- 何らかの指標を用いて使用する拡張後のデータを変化させて学習させる.
- 正規化辞書 (表記ゆれ対策) を作成する.
- 意味解析においては Juman は有効だがこのデータセットに対して有効かは疑問. unidic という現代語話し言葉コーパスなども使ってみる.

6 余談

先週末から, 今使っているノートパソコンの動作がとても不安定になってしまったので昨日, 買い換えました. なので, USAGI SERV の GPU の警告回りに気を配る余裕がなかったです.

7 今後の実験予定

- Data Augmentation の手法の改善案の模索.
- 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.

参考文献

[1] Chang M.-W. Lee K. Devlin, J. and K Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805*, 2018.