進捗報告

1 あらすじ

hottoSNS-BERT モデルが届いた.

2 進捗

- 未知語率測定

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 (京大 BERT･hottoSNS-BERT) を用いて 4 コマ漫画ストーリーデータセットについて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データオーギュメンテーション前後で算出した. 形態素解析には Juman++ を用いている.

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.

3.1 考察

3.1.1 語彙数

拡張前は京大 BERT の方が多かったが拡張後は hottoSNS-BERT の方が多かった. 理由としてはトークナイズ時の仕様の違いによるものだと考えられる.

- 京大 BERT は 形態素からサブワードへの変換は BPE を用いている
- hottoSNS-BERT は形態素解析を行わず SentencePiece を用いて直接テキストからサブワード化している

3.1.2 未知語率

前節の理由から例えば形態素「ただいま～」について
hottoSNS-BERT では

ただいま～

のままで全体として未知語として扱われているが京大 BERT では

ただ ##い ##ま ##～

と扱われ, その差が未知語率に大きく差をもたらしていると考えられる.

4 今後の実験予定

- Juman++ ではなく SentencePiece の事前学習済みモデルを用いて形態素解析を行う.
- セリフからの話者推定

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							
	京大 BERT	総単語数	311	331	315	316	316
		未知語率	0.026	0.021	0.019	0.022	0.025
	hottoSNS-BERT	総単語数	270	289	274	276	276
		未知語率	0.133	0.131	0.128	0.127	0.120
拡張後							
	京大 BERT	総単語数	2633	2728	2664	2705	2708
		未知語率	0.090	0.089	0.088	0.089	0.089
	hottoSNS-BERT	総単語数	3030	3209	3089	3154	3154
		未知語率	0.695	0.700	0.696	0.698	0.697