進捗報告

1 あらすじ

hottoSNS-BERT モデルが届いた.

2 進捗

- Pytorch モデルへの変換
- 未知語割合測定

3 hottoSNS-BERT

`UMR:\DataSet\Model`

内に TensorFlow 版と変換した Pytorch 版をそれぞれ置いている.

4 データセットに含まれる未知語割合

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.

4.1 考察

4.1.1 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

4.1.2 未知語割合

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, その差が未知語割合に大きく差をもたらしていると考えられる.

表 1: データセットに含まれる未知語割合

			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
		未知語割合	0.026	0.021	0.019	0.022	0.025
	hottoSNS-BERT	総単語数	270	289	274	276	276
		未知語割合	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
		未知語割合	0.090	0.089	0.088	0.089	0.089
	hottoSNS-BERT	総単語数	3030	3209	3089	3154	3154
		未知語割合	0.695	0.700	0.696	0.698	0.697

表 2: SentencePiece による形態素解析

	hottoSNS-BERT		ギャグ	少女	少年	青年	萌え系
拡張前							
	Juman++ 形態素化	総単語数	270	289	274	276	276
		未知語割合	0.133	0.131	0.128	0.127	0.120
	SentencePiece 形態素化	総単語数	318	342	327	332	329
		未知語割合	0.012	0.012	0.009	0.012	0.012
拡張後							
	Juman++ 形態素化	総単語数	3030	3209	3089	3154	3154
		未知語割合	0.695	0.700	0.696	0.698	0.697
	SentencePiece 形態素化	総単語数	2495	2597	2523	2568	2563
		未知語割合	0.036	0.035	0.035	0.035	0.035

4.2 SentencePiece による形態素解析

hottoSNS-BERT モデルのみについて, hottoSNS-BERT の分かち書きコーパス構築時に学習された SentencePiece の事前学習済みモデルを用いて形態素解析した場合についても未知語割合を算出した.

4.3 考察

4.3.1 未知語割合

Juman++ と比べて大幅に改善され, 拡張前後どちらの場合についても京大 BERT よりも未知語割合は下がった.

4.3.2 重要な未知語

['A', 'DVD', 'B', 'GPU']

4.3.3 データ拡張由来の未知語

['坐臥', '斯', '迚', '爾', '寔', '悉', '頗', '假', '遁', '顔', '佞', '兇', '譎', '奸譎', '瀟', '窈窕', '婉', '嬋', '悍', '裳', '艶', '閨', '嚊', '嬶', '庸', '灑', '滌', '潔', '臍', '浚', '婢', '嘱', '饒', '噺', '恤', '賦', '婬', '婬靡', '靡', '芍', '辯', '僉', '虞', '尤', '緻', '稟', '稟', '稟賦', '忽', '儕', '儷', '豁', '帙', '韋', '柢', '繙', '誦', '帛', '稍', '聊', '纔', '做', '搦', '疵瑕', '瑕疵', '疵', '昊', '穹', '稗', '躬', '輓', '靚', '扣', '撥', '斥', '洩', '擯斥', '兌', '壟', '渥', '乍', '軅', '孰', '已', '疚']

5 今後の実験予定

- 前期研究分を hottoSNS-BERT で再実験
- 二分割交差検証
- データセット不備への対処 (ラベルなしセリフをスキップしてデータから除外されていた)
- セリフからの話者推定