

1. 研究背景

1. Bidirectional Encoder Representations from Transformers (BERT) はパラメータが非常に多く、学習に膨大な時間が必要

2-1. BERTの事前学習において、finetuningするタスクに合わせたドメインのコーパスを利用することで性能が向上したという報告あり

2-2 ドメイン依存のテキストは大量には収集しにくい

「汎用的な小規模コーパス+ドメイン特有のコーパス」でBERT事前学習の効率化を目指す

2. 研究概要

1. 複合コーパスに基づくBERTモデルの構築

3. 大規模コーパスで学習したBERTの結果と比較

2. 特定のドメインに基づく固有表現認識 (NER)

タマネギとニンジン を 切って、
食材 食材 動作

学習時間を削減しつつ、モデルの性能を大規模コーパスで学習した際のそれとほぼ同等に維持することができるのか？

4. データセットの情報

データセット	行数	単語数	文字数
WikiText-JA	556,262	14,956,507	37,347,202
記事無作為抽出*	835,595	13,972,443	37,344,631
ゲーム解説コーパス	741,405	11,676,046	25,952,440
レシピコーパス	11,788,955	214,058,693	529,672,344

データセット	文数	単語数	固有表現数	固有表現クラス数
レシピコーパス				
学習	1,614	34,802	14,058	
開発	66	1,326	527	21
テスト	67	1,345	597	
ゲーム解説コーパス				
学習	1,178	23,368	8,432	
開発	235	3,369	1,266	8
テスト	225	3,375	1,540	

表1 BERTの事前学習に用いたコーパス諸元

表2 NERに用いたコーパス諸元

* 記事無作為抽出: 文字数に関してWikiText-JAと同サイズとなるように、日本語版Wikipediaから無作為に記事を抽出したデータセット

3. WikiText-JAとは？

1. データ収集

2. クレンジング処理

WikiText-JA

日本語版Wikipediaの様々なジャンルの記事が収録された汎用的な小規模コーパス

文法的な正確さを保った長文を数多く含み、かつ文章間で表記のゆれが抑えられていることが期待される

5. 実験条件

1. BERT事前学習時

2. NER実施時

* Wikipedia全体によるBERTのみ30 epochで学習したモデルを使用

6. 実験結果

データセット	レシピNER				事前学習にかかった時間	データセット	ゲーム解説NER				事前学習にかかった時間
	Prec. (%)	Recall (%)	F-meas.				Prec. (%)	Recall (%)	F-meas.		
WixiText-JA	87.4	89.9	88.6		32 hours	WixiText-JA	85.1	80.5	82.7		32 hours
レシピコーパス	89.9	92.2	91.0		31 hours	ゲーム解説コーパス	85.7	78.9	82.2		32 hours
Wikipedia全体	90.6	92.2	91.4		38 days	Wikipedia全体	87.1	81.9	84.4		38 days
記事無作為抽出*+レシピ	90.3	92.0	91.2		31 hours	記事無作為抽出*+ゲーム解説	86.3	80.4	83.2		31 hours
WikiText-JA+レシピ	90.7	92.5	91.6		32 hours	WikiText-JA+ゲーム解説	86.7	82.6	84.6		31 hours

WikiText-JAとドメイン特有のコーパスで事前学習したBERTは

1. 他のデータセットで事前学習したいずれのBERTよりも高い精度を示した

2. 短い学習時間で、Wikipedia全体で学習したBERTと同等以上の性能を獲得した

7. 結論

WikiText-JAとタスクの対象ドメインに関わるコーパスを組み合わせることで、学習時間を削減しつつもモデルの性能を大規模コーパスで学習した際のそれとほぼ同等に維持できた

8. 今後の展望

1. 学習が進んでいくにつれての、複合コーパスで構築したBERTと日本語版Wikipedia全体で構築したBERTの精度変化の比較

2. BERT pretrainingにおけるドメイン特有なコーパスの影響の調査

各データセットに基づくBERTの性能を調査

WikiText-JA : 100 % W: 75 % W: 50 %
+ + +
レシピ : 0 % レ: 25 % レ: 50 %