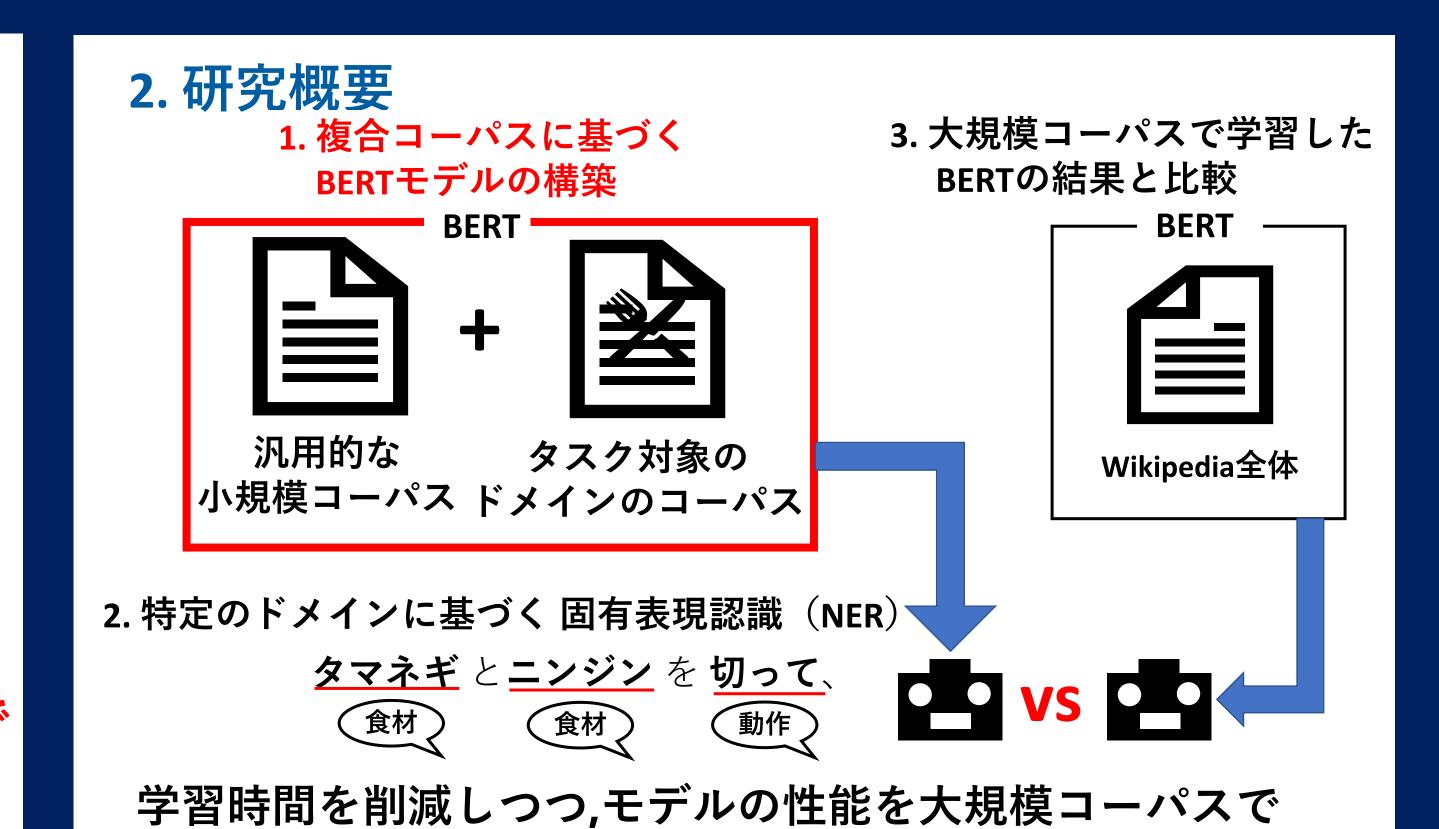
WikiText-JA構築によるBERT事前学習の効率化

小川晃1),友利凉1),亀甲博貴2),森信介2) 1京都大学大学院情報学研究科,2京都大学学術情報メディアセンター

1. 研究背景

- 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



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

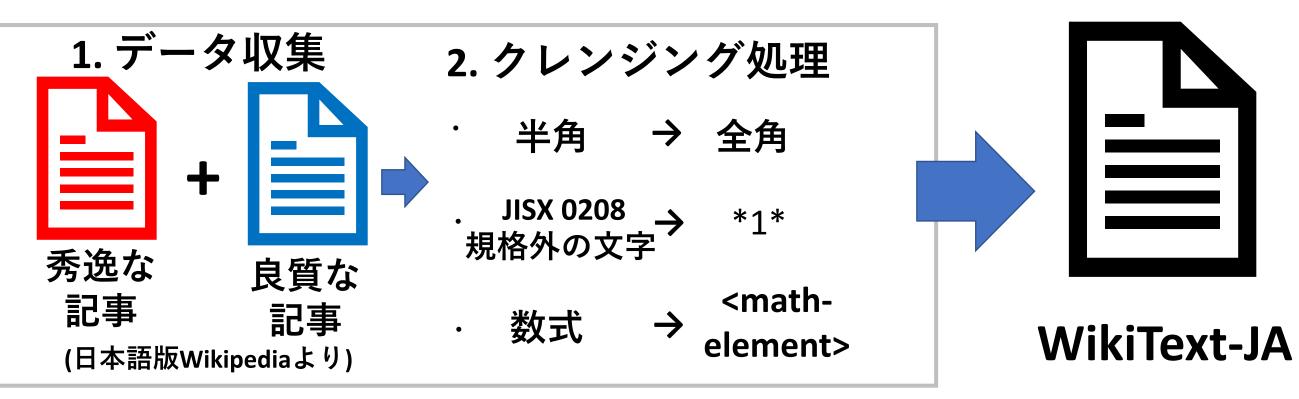
データセット	行数	単語数	文字数	データセット	文数	単語数	固有表現数	固有表現クラス数
				レシピコーパス				
WikiText-JA	556,262	14,956,507	37,347,202	学習	1,614	34,802	14,058	
				開発	66	1,326	527	21
記事無作為抽出*	835,595	13,972,443	37,344,631	テスト	67	1,345	597	
				ゲーム解説コーパス	ス			
ゲーム解説コーパス	741,405	11,676,046	25,952,440	学習	1,178	23,368	8,432	
レシピコーパス	11,788,955	214,058,693	529,672,344	開発	235	3,369	1,266	8
				テスト	225	3,375	1,540	
±4 DDD=	+ + + + + = = = = = = = = = = = = = = =				0 NED4-			

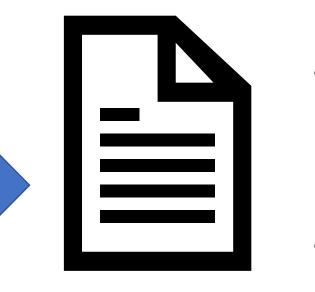
表1 BERTの事前学習に用いたコーパス諸元

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?





- 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
- 入力の最大文長: 128
- ・バッチサイズ 32
- 語彙サイズ: 32000
- 学習時のステップ数: 200,000*

- 2. NER実施時
- バッチサイズ: 32 (訓練), 8 (開発)
- 訓練時のエポック数:4
- ・訓練:開発:テスト=8:1:1

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

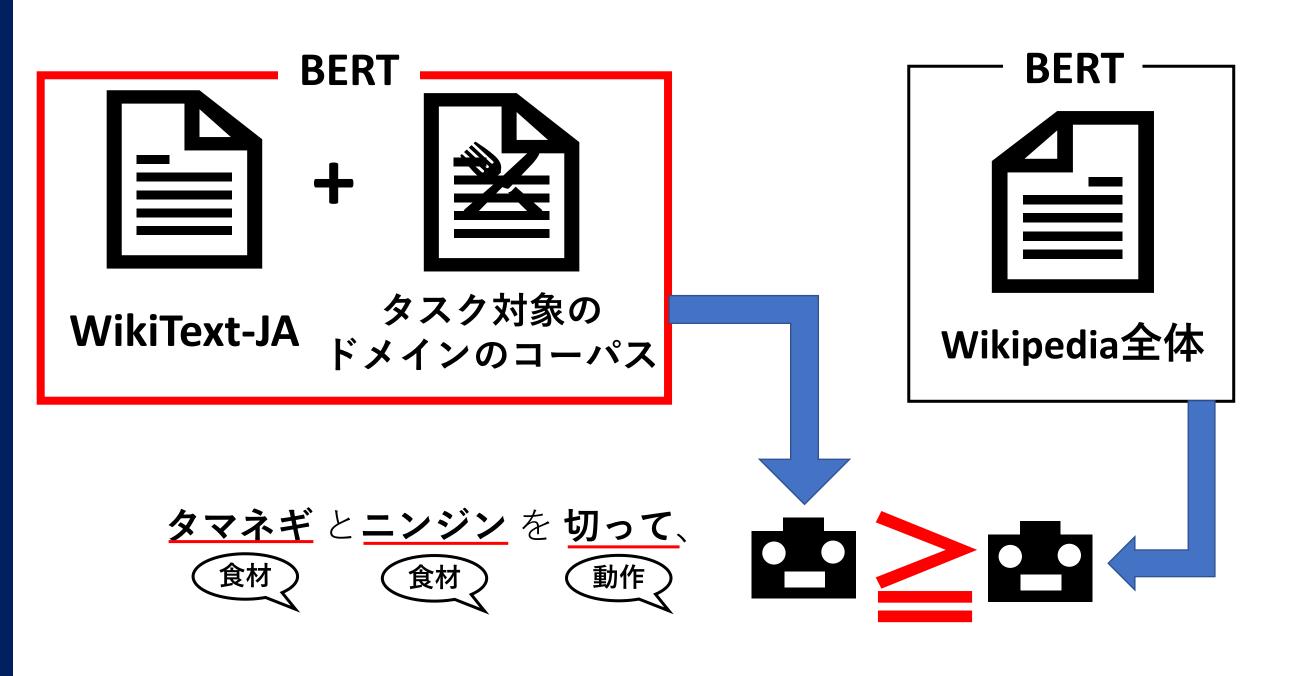
6. 実験結果

	レシピNER					ゲーム解説NER				
データセット	Prec. (%)	Recall (%)	F-meas.	事前学習に かかった 時間	データセット	Prec. (%)	Recall (%)	F-meas.	事前学習に かかった 時間	
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

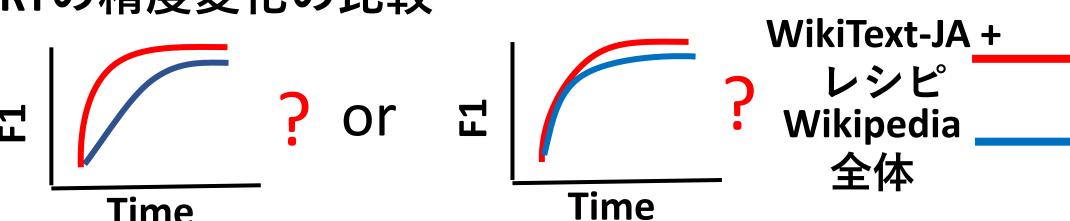
7. 結論



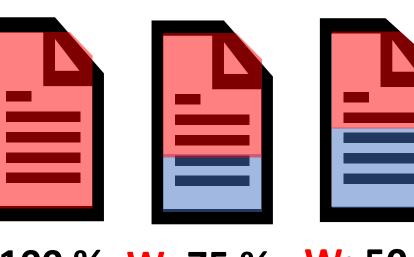
WikiText-JAとタスクの対象ドメインに関わるコーパスを 組み合わせることで、学習時間を削減しつつもモデルの性能を 大規模コーパスで学習した際のそれとほぼ同等に維持できた

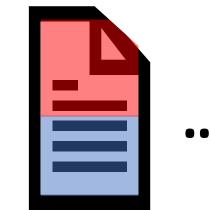
8. 今後の展望

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%