論　文　内　容　の　要　旨
Synopsis of Thesis

氏　　名 Name	和田　聖哉
論文題名 Title	Oversampling effect in pretraining for bidirectional encoder representations from transformers (BERT) to localize medical BERT and enhance biomedical BERT （医学BERTモデルのローカライゼーションと生物医学BERTモデルの強化：オーバーサンプリングによる事前学習のアプローチ）

論文内容の要旨

〔目　的(Purpose)〕

生テキストから直接事前学習を行うことで生成される大規模ニューラル言語モデルを利用して別の目的タスクを解く戦略（転移学習、transfer learning）は、現代の自然言語処理において標準的なアプローチである。2018年、transformer技術をベースに構築された言語モデル（Bidirectional Encoder Representations from Transformers：BERT）は、自然言語処理の精度を大幅に向上させた。生物医学分野でも同様の進展が見られるが、言語資源に乏しい専門領域では、その分野に特化した言語モデルを構築すること自体が困難である。我々は、特定領域に焦点を当てたコーパスをオーバーサンプリングし、これを大規模コーパスと組み合わせることで、バランスの取れた事前学習が可能になり、この課題を克服できると仮定した。

〔方法ならびに成績(Methods/Results)〕

特定専門領域のコーパスにオーバーサンプリングを適用し、バランスよく事前学習して言語モデルを開発し、そのパフォーマンスを既存手法モデルと比較した。我々は、英語と日本語のモデルについて、それぞれbiomedical language understanding evaluation (BLUE)ベンチマークと複数の日本語自然言語処理タスクを用いて性能評価を行い、提案手法の有効性を検証した。

まず英語の小規模生物医学コーパスと一般領域コーパスを使用して生物医学BERTモデルを構築したところ、提案手法のモデルはBLUEベンチマークの10個のタスクのうち、6つのタスクで統計的有意差を示した。また、小規模生物医学コーパスのサイズ毎にモデルを構築して既存手法との性能比較を行ったところ、全ての条件で提案手法によるモデルの性能は有意に高値であり、提案手法の汎用性を確認した。

英語モデルの成功に続き、日本語でも同様に小規模医学コーパスと一般領域コーパスを使用してBERTモデルを構築した。4種類の日本語タスクのうち、3種類で提案手法のモデルが最も高い精度を示した。そのうち2種類のタスクでは比較した他のモデルよりも統計的な有意差を認め、言語が異なっても有用な手法であることを確認した。

これらの実験結果に基づき、さらなる効果量の分析を行うために追加の実験を実施した。最初に、PubMedに収録されている記事のうち、ヒトの疾患に関連するものとそれ以外の医学記事とを区別し、提案手法を適用し強化した医学モデルを構築した。このモデルを、それらのコーパスを区別せずにPubMedの全記事を使用して従来手法で構築したモデルと比較したところ、BLUEベンチマークでは生物医学スコアが0.49ポイント、臨床スコアが0.34ポイント向上した（$p<0.05$）。さらに効果量を評価するためにbiomedical BlueBERTとclinical BlueBERT、PubMedBERTとKeBioLMの2組の比較を行った。どちらの場合も臨床スコアは上昇したものの、生物医学スコアは低下する傾向を認めた。我々のモデルはPubMed記事のみを用いるという制約のもとで構築された。これらの結果は、臨床医学コーパスに提案手法を適用することで、既存のモデルを上回る性能を持つ言語モデルの構築も可能であることを示唆している。

〔総　括(Conclusion)〕

目的タスクに適したコーパスから派生したインスタンスをオーバーサンプリングしてバランス良く事前学習することで、高性能なBERTモデルを構築できることを示せた。本研究は、特定の専門領域でも、適切なコーパス選択と事前学習の工夫で高性能なBERTモデルが構築可能であることを実証した。このアプローチは、言語資源が限られている他の領域においても有効である可能性があり、さらなる研究の基盤を提供する。

論文審査の結果の要旨及び担当者

（申請者氏名）		和田　聖哉	
	（職）	氏　　名	
論文審査担当者	主　査　大阪大学教授	武田 理宏	署名
	副　査　大阪大学教授	川崎 良	署名
	副　査　大阪大学教授	服部 純一	署名

論文審査の結果の要旨

この論文では、大量の生テキストから深層学習を用いて言語モデルを構築する方法、特にBERT（Bidirectional Encoder Representations from Transformers）モデルを中心に、その情報抽出性能の改善について論じている。生物医学分野で応用事例が多く報告されているものの、高品質な大規模コーパスが不足している領域では、精度の高い言語モデルの構築が困難であった。本研究では、英語の生物医学BERTモデル、日本語の医学BERTモデル、およびPubMed抄録を用いて事前学習し強化された生物医学BERTモデルという3つの検証を通じて、実効性、頑健性、汎用性を実証し、提案手法が既存の手法よりも優れていることを示した。これにより、特に日本語の医療情報抽出や言語モデルの開発を通じて、高性能な言語モデルの開発に大きな貢献が期待される。

この研究はその実用性と応用性から、博士（医学）の学位授与に値すると考えられる。