別添４

厚生労働科学研究費補助金（政策科学総合研究事業）
（臨床研究等 ICT 基盤構築・人工知能実装研究事業）
分担研究報告書

課題名　　：新薬創出を加速する症例データベースの構築･拡充/創薬ターゲット推定アルゴリズムの開発

研究分担者名：黒橋 禎夫
国立大学法人京都大学大学院 情報学研究科 知能情報学専攻 教授

研究要旨
医療分野における臨床テキスト、患者テキストの解析・構造化のための言語・知識処理基盤を構築し、特発性肺線維症および肺がんの創薬標的予測に資するテキスト構造化を実現する。今年度は、以下２つの研究を実施した。

- 大規模なラベル無しの医学テキストを活用する知識抽出
- 診療録のアノテーション拡大の効果と深層知識抽出モデルの精度の考察

A．研究目的

本プロジェクトではこれまで、アノテーション付き医療テキストコーパスと、これを学習用データとして活用した医療エンティティ・属性・関係認識システムを構築し、医療テキストの構造解析・情報抽出に取り組んできた。

深層学習モデルによって正確な医療テキスト構造化を実現するためには高品質な医療アノテーション・コーパスの構築が不可欠であるが、人手アノテーションを行うのはコストと時間がかかる。本研究では、人手アノテーションへの過度な依存を防ぎつつ、深層モデルの学習にデータ拡大の必要性を解明するために、以下の2つの研究を実施する。

① 大規模なラベル無しの医学テキストを活用する知識抽出
② 診療録のアノテーション拡大の効果と深層知識抽出モデルの考察

B．研究方法

① 大規模なラベル無しの医学テキストを活用する知識抽出：

現実において、医療データの量が不足しており、関係認識の精度が十分でない。Distant Supervisionを活用して、大量のラベルなしテキストから、人手コーパス中のエンティティペアを含む文を抽出し、半教師付きデータを事前学習する。図1のように、少数のラベル付きデータを活用し、半教師付きデータの信頼性を推定し、ノイズ除去を行うWeighted Contrastive Learning（WCL）対照学習方法を提案した。本成果WCL-REは自然言語処理分野のトップ国際会議であるEACL2023に採択され、2023年5月に発表を行う予定である。

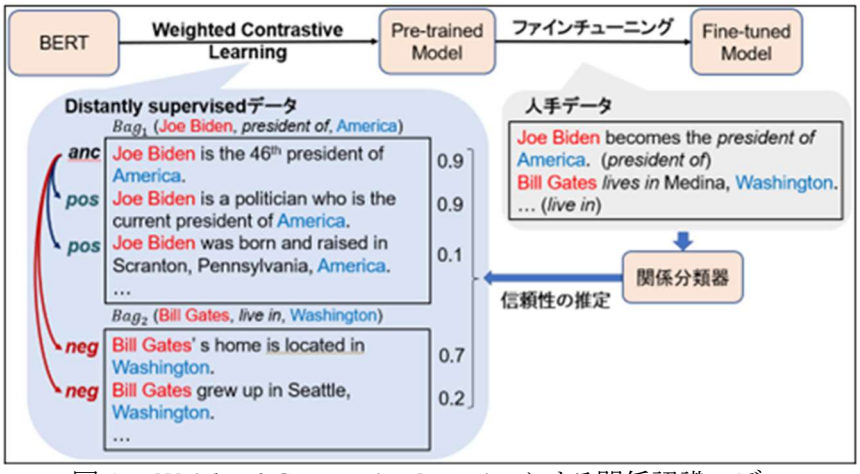

図１：Weighted Contrastive Learning による関係認識モデル

データを活用する研究では、事前学習をせずに半教師データを利用する方法も考える。半教師データ中の例との意味的類似度を推定して、最近傍事例（kNN: k-Nearest Neighbor)を用いたデータ利用効率の向上手法を提案した。図 2 では、予測例が半教師データから最近傍事例を検索し、ラベルの重みを再構成し、モデルが予測確率を修正する。本成果 kNN-RE は 2022 年 12 月にトップ国際会議である EMNLP2022 で発表した。

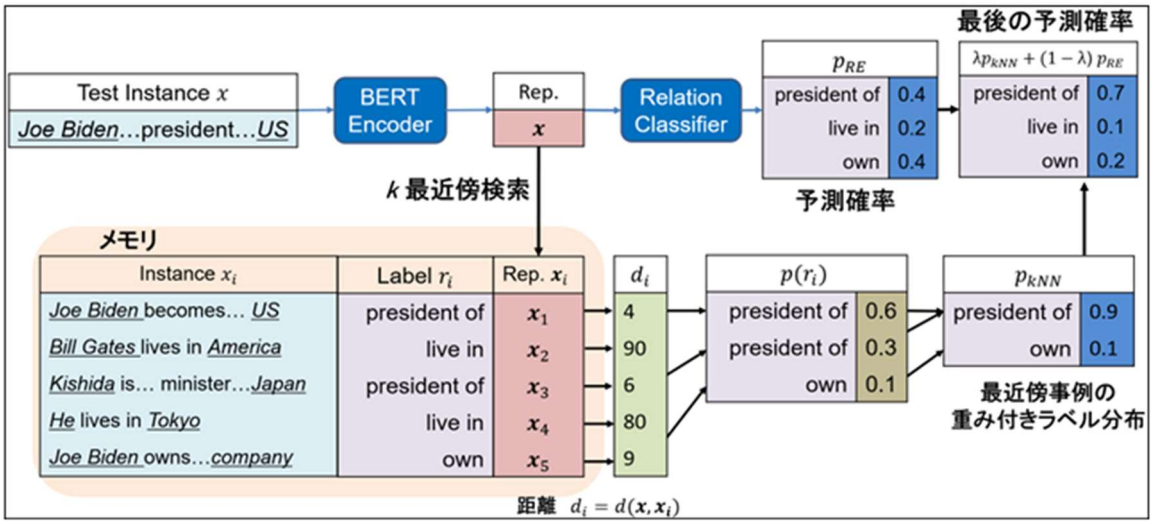

図 2：最近傍事例を用いた関係認識モデル

② 診療録のアノテーション拡大の効果と深層知識抽出モデルの考察：
昨年度、各 IPF 患者に対する複数の診療歴を含む新規の 1,500 件の診療録を入手し、医療事象と時間表現間の時間関係を含む PRISM 臨床医学テキストアノテーション仕様に従って、人手アノテーションを行った。今年度、アノテーションの修正（特に「治療」と「病変・症状」）を実施した。BERT を用いた Joint エンティティ・モダリティ・関係抽出モデルを利用し、1,500 件の診療録データでの学習と既存の 371 件の性能を比較した。

（倫理面への配慮）
本研究は、医薬基盤・健康・栄養研究所において倫理審査、承認を得た後、人を対象とする医学系研究に関する倫理指針に従って実施している。

C．研究結果
① 大規模なラベル無しの医学テキストを活用する知識抽出：
図 3 において、WCL-RE 手法は半教師データで事前学習されていない Baseline を大幅に上回る。特に、学習データの 25%だけを使用する設定では、提案手法は F 値が 1.64~21. 41 ポイント向上した。ノイズ除去を行う WCL-RE は、他の対照学習手法（CIL、RECN)と比較して、さらに優れた性能を示した。

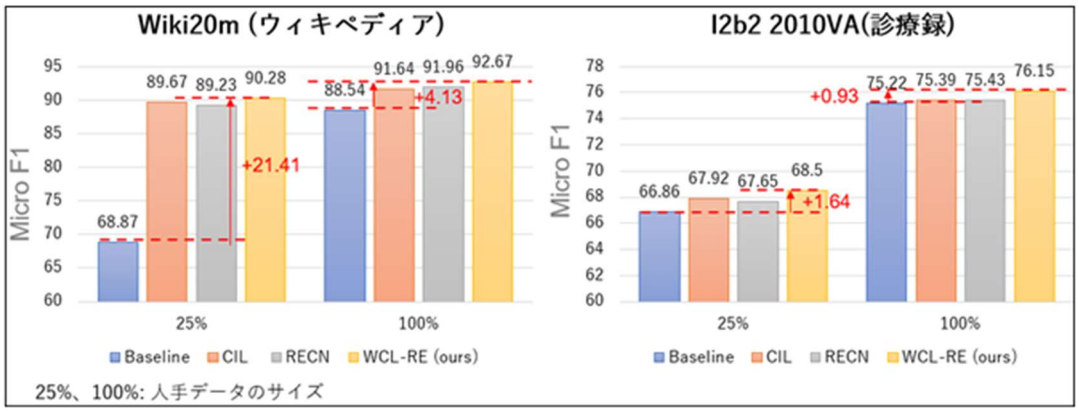

図 3：事前学習言語モデルによる医療エンティティ表現の置換

kNN-RE は全てのデータセットで Baseline を超え、3 つのデータセット（ACE05, Wiki80, SciERC）で SOTA を達成した。図 4 では、半教師付きデータが利用可能な Wiki80 や i2b2 2010VA で、半教師付きデータメモ

リによるさらなる F 値向上が確認できる。

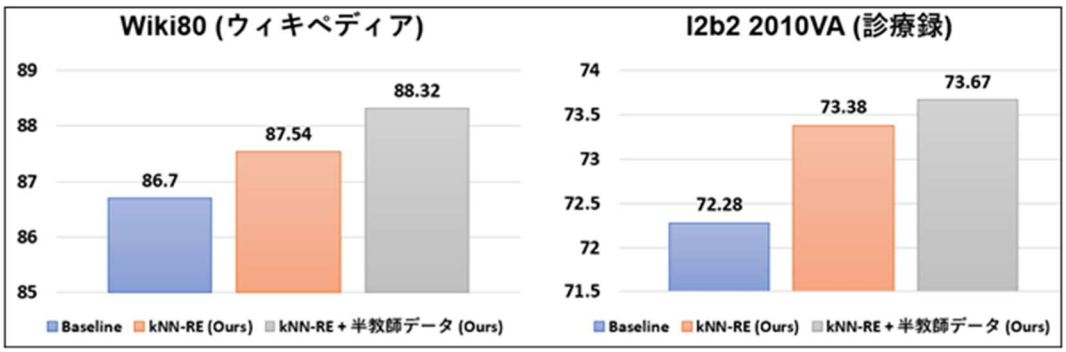

図 4：kNN-RE 関係抽出の F 値

② 診療録のアノテーション拡大の効果と深層知識抽出モデルの考察：

1,500 件診療録を用いて知識抽出実験を行った F 値結果は 92.91 (実体), 88.13 (モダリティ), 82.83 (関係) を達成した。この結果は、従来の 371 件の結果よりも大幅に改善されている。

D．考察

① 大規模なラベル無しの医学テキストを活用する知識抽出：

図 5 において、Low-Resource 設定で kNN-RE は非常に優れた性能を示す。1%学習データを使う場合、半教師データをメモリとすることで大幅な改善（+42.35 F 値）を達成している。

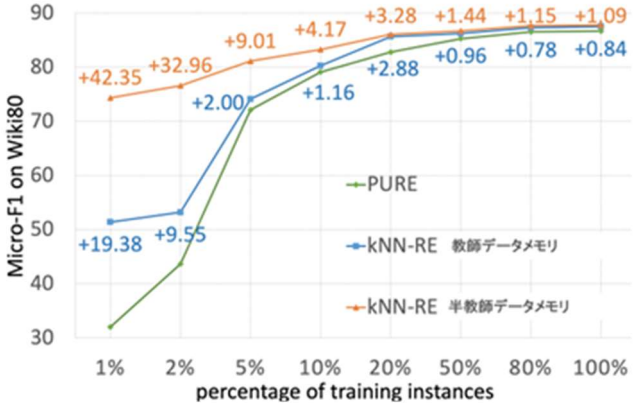

図 5：Low-Resource 設定の結果

② 診療録のアノテーション拡大の効果と深層知識抽出モデルの考察：

図 6 において、学習データ量の増加に伴い、三つのタスク（特に関係抽出）での F 値が大幅に向上した。

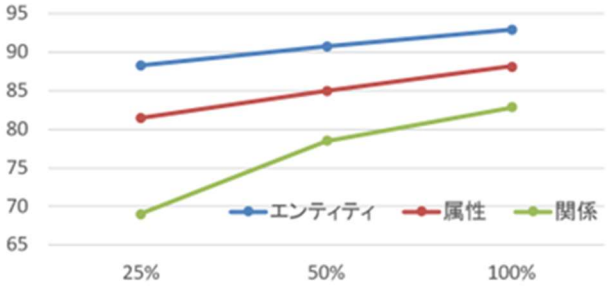

図 6：1,500 件 IPF 診療録の学習曲線

E．結論

① 大規模なラベル無しの医学テキストを活用する知識抽出：

WCL-RE と kNN-RE 手法はどちらも、半教師データを活用して関係抽出タスクの改善を図り、複数のデータセットで SOTA を達成した。kNN-RE は Low-Resource 設定において非常に優れた性能を示しており、将来的には few-shot、zero-shot 設定でも利用可能である。

② 診療録のアノテーション拡大の効果と深層知識抽出モデルの考察：
1,500 件の診療録で学習されたモデルは、既存の 317 件よりも大幅に優れている。この結果は、深層学習モデルがデータ hungry であり、人手アノテーションが利用可能になる場合にはさらなる改善が期待できることを示唆する。

F．研究発表
1．論文発表
なし

2．学会発表
1) Zhen Wan, Qianying Liu, Zhuoyuan Mao, Fei Cheng, Sadao Kurohashi and Jiwei Li: Rescue Implicit and Long-tail Cases: Nearest Neighbor Relation Extraction, Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022), Abu Dhabi, (2022.12).
2) Zhen Wan, Fei Cheng, Qianying Liu, Zhuoyuan Mao, Haiyue Song and Sadao Kurohashi: Relation Extraction with Weighted Contrastive Pre-training on Distant Supervision, Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2023): Findings Volume, Dubrovnik, Croatia, (2023.5) (accepted).

G．知的財産権の出願・登録状況（予定を含む。）
1．特許取得
なし

2．実用新案登録
なし

3．その他
なし