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、特発性肺線維症及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 つの問題設定で研究を実施した。

- 「重要事象に焦点を当てた症例報告の軽量なグラフ構造要約」を提案した。グラフ構造要約を弱教師とする症例報告の自動構造化により、十分医師の参考となるレベルの解析精度（F値69.8）を達成した。
- 「日本語医療テキストの関係抽出におけるドメイン適応」を提案した。日本語医療テキストにおいて、事前学習言語モデルによる単語置換に基づくドメイン適応手法により肺がんとIPFドメイン間での精度向上を確認した。

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。

しかし、肺がん、特発性肺線維症（IPF）以外のターゲット疾患の医療テキスト解析の実現するためには、現在のアプローチではまずターゲット疾患のテキストの大規模コーパスを構築する必要がある。医療テキスト中のすべての言語現象にアノテーションした、大規模かつ汎用的なコーパスの構築にかかるコストは大きく、ターゲット疾患ごとに同様の手法を適用することは現実的ではない。

そこで、ターゲット疾患の大規模タグ付きコーパスが存在しないという現実的な問題設定を考え、知識を活用して医療テキストの解析を実現する研究に取り組む。具体的には以下の2つのアプローチを取る。

① 重要事象に焦点を当てた症例報告の軽量なグラフ構造要約
② 日本語医療テキストの関係抽出におけるドメイン適応

Ｂ．研究方法

① 重要事象に焦点を当てた症例報告の軽量なグラフ構造要約：

既存の網羅的な医療情報アノテーションを行うかわりに、症例報告における重要事象を構造的に要約し、症例報告の効率的な検索・利活用を実現することが目標である。図１に症例報告の一例とそのグラフ構造要約を示す。症例報告における重要な事象（病名、臓器部位など）は「病気とその場所」、「検査とその結果」などの関係で結ばれ、グラフ構造をなす。

タイトル: 無症候性の虚血性腸炎を認めた全身性エリテマトーデスの 1 例

症例: 65 歳、女性。主訴:発熱と体重減少。現病歴: 2000 年関節痛が出現し、近医で抗 RNP 抗体単独陽性から混合性結合組織病と診断されステロイド内服加療を行っていた。2007 年 5 月発熱、全身倦怠感、体重減少のため当院に入院した。リンパ球減少、関節炎、抗核抗体陽性、抗 DNA 抗体陽性から全身性エリテマトーデス(SLE)と診断した。腹部症状は認めなかったが、大腸内視鏡検査で多発性直腸潰瘍を認め、病理組織で虚血性腸炎と診断した。全身性エリテマトーデスによる血管炎が原因と考え、シクロフォスファミド点滴静注療法(IVCY)を行い潰瘍病変の改善を認めた。

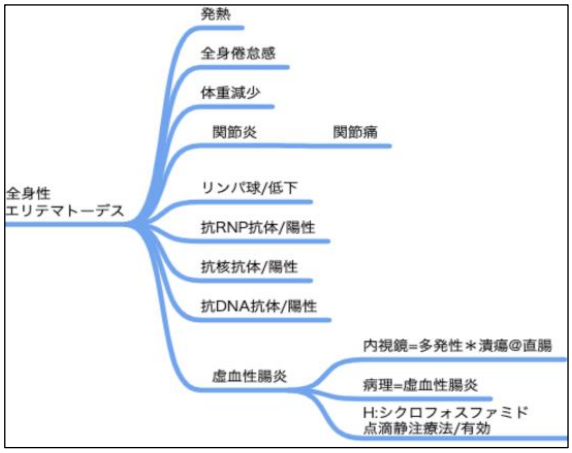

図 1 ：症例報告とグラフ構造要約の例

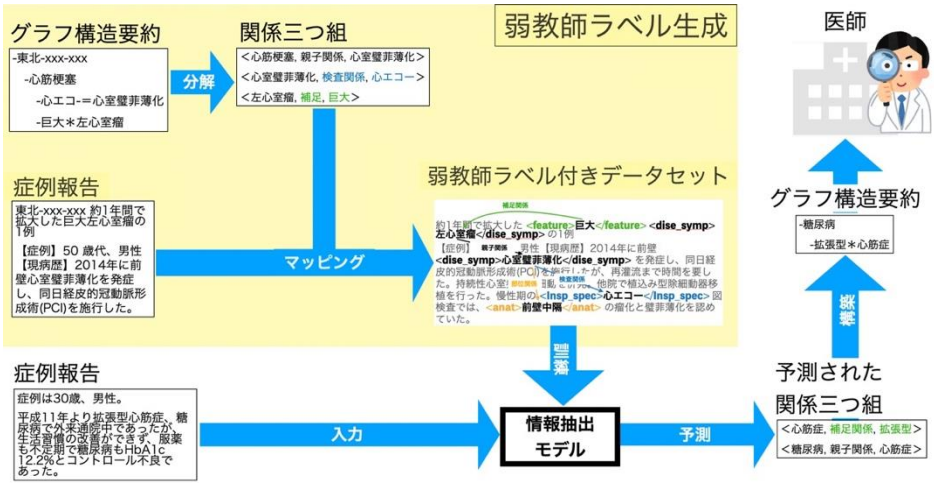

図 2：弱教師学習に基づく症例報告の構造的要約の流れ

本研究では、症例報告からグラフ構造要約を自動生成することに取り組む。構造要約中の各事象を症例報告中の言及箇所に対応づけることで情報抽出の弱教師ラベル付きデータを構築し、深層学習言語モデル BERT により症例報告中の重要事象や事象間の関係を予測することでグラフ構造要約を生成する。提案手法の流れを図 2 に示す。

② 日本語医療テキストの関係抽出におけるドメイン適応：

日本語医療アノテーションデータを増やすことは二つの理由で困難である。まず、医療テキストは個人情報を含んでいることが多く、データを公開するには匿名性の確保など慎重な扱いが必要となる。次に、アノテーションに専門的な知識が必要であり、多大なコストがかかる。そこで、すでに存在するアノテーション付き医療テキストを利用し、PRISM 医療関係抽出データで肺がんと IPF の二つのドメイン間の適応実験を行う。本研究では関係抽出タスクに取り組む。関係抽出は例えば、「両側頸部に小リンパ節を散見する」というテキストから(両側頸部, region, 小リンパ節)、(散見, feature, 小リンパ節)のような関係を抽出するタスクである。本研究では、大規模な日本語の医療テキストで事前学習言語モデル（UTH BERT）によってソースデータ中の医療エンティティ表現を置換するドメイン適応法を提案する。提案手法の流れを図 3 に示す。

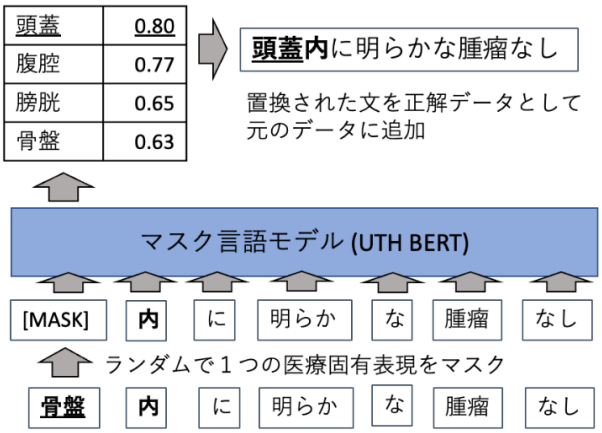

図 3：事前学習言語モデルによる医療エンティティ表現の置換

（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 重要事象に焦点を当てた症例報告の軽量なグラフ構造要約：

実験では約 15,000 件のグラフ構造要約データもとで日本語情報抽出モデルを学習し、精度として F 値 69.8 を達成した。実験結果を表 1 に示す。英語の医療情報抽出のデータセットである i2b2/VA 2010 では、現時点で BERT ベースの性能は F 値で 60 程度、日本語医療ドメインの関係抽出の教師付きデータセットにおける BERT ベースモデルの性能は F 値で 76 程度である。データが異なるため直接の比較はできないが、本研究が弱教師学習であることを考慮すると、全体での精度 F 値 69.8 は良い結果だと言える。

表 1：関係抽出の精度

適合率	再現率	F 値
80.5	63.5	69.8

② 日本語医療テキストの関係抽出におけるドメイン適応：

提案したドメイン適応手法によってソースドメインで学習を行い、ターゲットドメインで評価する実験を行った。実験結果を図 4 に示す。肺がんドメインで学習し、IPF ドメインでテストした場合、ドメイン適応によって F1 が 0.9 ポイント向上した。逆に、IPF で学習し、肺がんでテストした場合は F1 が 8.0 ポイント向上した。

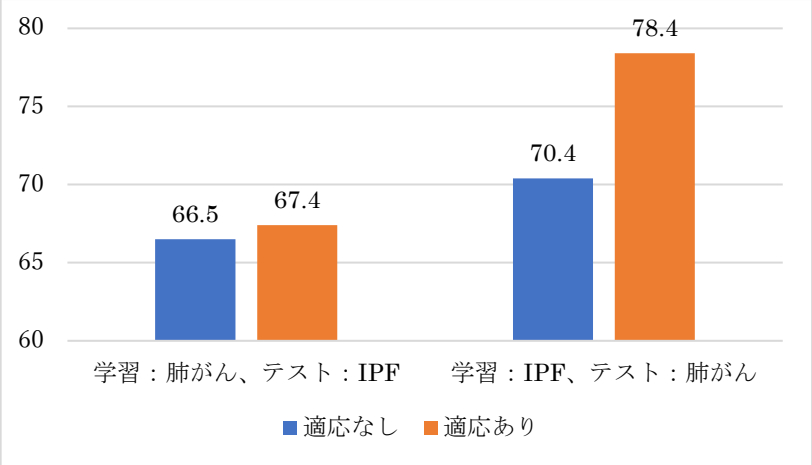

図 4：ドメイン適応による関係抽出精度の向上（F 値）

D．考察

① 重要事象に焦点を当てた症例報告の軽量なグラフ構造要約：

実験結果に対して、医師の定性的分析により、本システムの予測結果は医師がグラフ構造要約を作成する際に十分に参考となるレベルであり、要約基準の一貫性の向上にも資するものと考えられる、との評価を得

た。医師による人手の評価を通してモデルの精度を定量的に評価することは今後の課題である。
② 日本語医療テキストの関係抽出におけるドメイン適応：
実験結果において、学習データに肺がんを用いた場合と IPF を用いた場合とで精度の向上に差がみられる。これらの差は、UTH BERT の事前学習データの影響だと考えられる。事前学習データのドメインには分布に偏りがある。UTH BERT によって置き換えられる単語は事前学習データに比較的よく現れる肺がん関連の単語であることが多く、結果として肺がんのテストに対する精度が大きく向上したと考えられる。

Ｅ．結論
①　重要事象に焦点を当てた症例報告の軽量なグラフ構造要約：
本研究では、グラフ構造要約を関係抽出の弱教師として症例報告を構造化するフレームワークを提案した。関係三つ組抽出の精度はＦ値で 69.8 であった。今後は技術的改善を行うとともに、本研究をグラフ構造要約データの拡張に活用したい。
② 日本語医療テキストの関係抽出におけるドメイン適応：
本研究では、日本語医療テキストの関係抽出におけるデータ拡張を用いたドメイン適応手法を提案した。大規模な日本語医療テキストで事前学習された言語モデルを使うことで、置き換え単語の多様性が増し、精度が向上することを示した。

Ｆ．研究発表
１．論文発表
なし

２．学会発表
1) Zhen Wan, Fei Cheng, Zhuoyuan Mao, Qianying Liu, Haiyue Song, Sadao Kurohashi: Improving Medical Relation Extraction with Distantly Supervised Pre-training, 言語処理学会 第 28 回年次大会, 浜松, (2022.3.14).
2) 尾崎 立一, 清丸 寛一, Cheng Fei, 黒橋 禎夫, 佐藤 寿彦, 永井 良三: 弱教師学習に基づく症例報告の構造的要約, 第 26 回日本医療情報学会春季学術大会, 岡山, (2022.6.30)（採録決定）

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