弱教師学習に基づく症例報告の構造的要約

尾崎 立一[*1], 清丸 寛一[*1] , Cheng Fei[*1], 黒橋 禎夫[*1], 佐藤 寿彦[*2], 永井 良三[*3]
[*1] 京都大学大学院情報学研究科, [*2] 株式会社プレシジョン, [*3] 自治医科大学

Structured Summarization of Case Reports Based on Weakly Supervised Learning

Ryuichi Ozaki[*1], Hirokazu Kiyomaru[*1], Fei Cheng[*1], Sadao Kurohashi[*1]
Hisahiko Sato[*2], Ryozo Nagai[*3]
[*1] Graduate School of Informatics, Kyoto University
[*2] Precision Co., Ltd., [*3] Jichi Medical University

抄録:科学などの専門化・細分化が進んだ医学分野では，過去の症例を参照する必要がたびたび生じる．本稿の共著者である永井は症例報告の効率的な検索を実現するため，症例報告に対する要約（以下グラフ構造要約）の基準を定義し，1 万 5 千件規模の症例報告の要約を手作業で作成、さらに 3 万 5 千語の検索用辞書を整備した．本研究では，このグラフ構造要約作成の作業コストを削減し一貫性を確保するため，自然言語処理により症例報告からグラフ構造要約を自動生成する手法を示す．具体的には，グラフ構造要約の各要素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データを構築し，機械学習により症例報告中の重要要素や関係を予測することでグラフ構造要約を生成する．実験の結果，F 値 69.8 でグラフ中の関係を抽出することに成功した．

キーワード　診断支援システム, グラフ構造要約, 自然言語処理, 情報抽出

1.　はじめに

医学，特に内科学は専門化と細分化が進み，一人の臨床医が患者さんに対して総合的視点で診断を行うことが極めて難しい状況が生まれつつある．この問題は，過去の症例を蓄積・構造化し効率的に検索できるようになれば，かなり緩和・解決するものと考えられる．本稿の共著者である永井は症例報告に対する要約（本稿ではグラフ構造要約と呼ぶ）の基準を定義し，さらに永井自身が数千件規模の症例報告の要約を手作業で作成し，ウェブブラウザでの検索機能を備えた J-CaseMap を開発している[1]．この活動・データベースを今後拡大していく上での課題は，医師によるグラフ構造要約の作業コストの軽減と一貫性の確保であった．

本稿では，自然言語処理技術によって，症例報告からグラフ構造要約を自動生成する手法を示す．約 15,100 件のグラフ構造要約を弱教師とする手法により，F 値 69.8 でグラフ中の関係を抽出することに成功した．この結果は，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．

2.　症例報告のグラフ構造要約

1）グラフ構造要約の例・説明

図 1 に症例報告の一例とそのグラフ構造要約を示す．グラフ構造要約は，ルートにその症例の主な病名を持つ木構造で，ルート以外のノードには病名，所

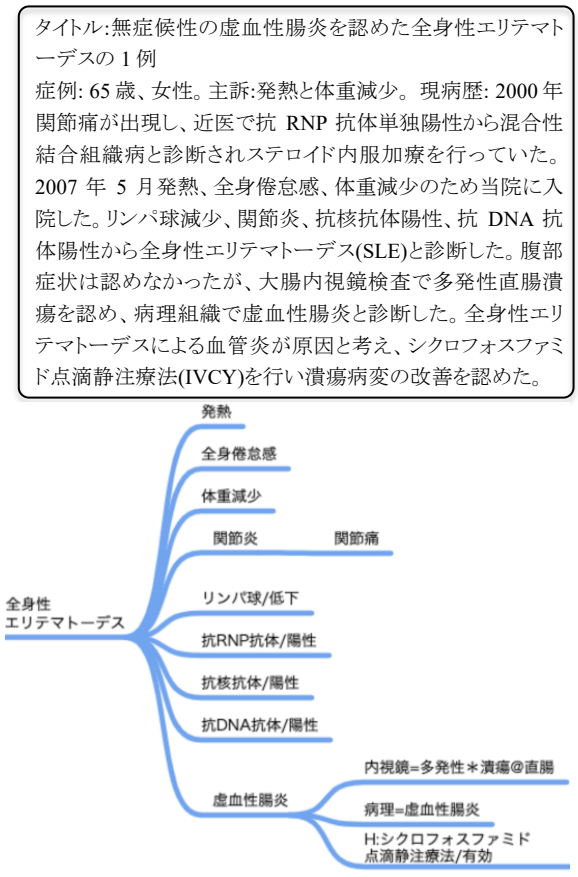

図 1　症例報告とグラフ構造要約の例．“＊”, “＝”は左から係り補足情報，検査・検体名を，“@”, “/”は右から係り解剖部位，正負を持つ情報を表す．

見などを持つ．親子関係はストーリーと呼ばれる緩いつながりを表す．各ノードは医療エンティティと，図 1 の注釈に示す 4 種類の特殊な記号からなり，それ自体が構造を持つ．

2）自然言語処理から見たグラフ構造要約

自然言語処理の視点からは，グラフ構造要約の各要素は症例報告からの情報抽出（Information Extraction: IE）の結果であると見ることができる．IE の基本的な手法は，テキスト自体にアノテーション（付加情報）を与え，そこから機械学習を用いてテキスト中の重要要素や関係を抽出する．しかし，本課題で利用できるのはグラフ構造要約であり，アノテーションされたテキストではない．

そこで，本研究ではこのグラフ構造要約中のエンティティを症例報告中の言及箇所に対応づけることで IE の弱教師として利用し，IE モデルを訓練して，症例報告からグラフ構造要約を自動構築するシステムを提案する．

3. 方法

1）弱教師生成

まずグラフ構造要約中のエンティティを症例報告中の対応する言及箇所にマッピングする．具体的にはエンティティを同義語展開し，3 文字以下の文字の挿入・削除を許したルールベースの文字列マッチングによって行われる．

次にグラフ構造要約を分解し，記号と木構造内の親子関係から以下の 4 種類の関係 3 つ組と 1 種類のモダリティに分解する．

- 親子関係を持つノード間の病名・症状間に貼られる「**親子関係**」
- 病名・症状とその検査・検体名の間に貼られる「**検査関係**」
- 病名・症状とその解剖部位の間に貼られる「**部位関係**」
- 病名・症状や検査・検体名、解剖部位とその補足情報の間に貼られる「**補足関係**」
- “/”で表される病名・症状の**正負情報のモダリティ**

こうして得られた関係・モダリティの情報を各エンティティに対応する言及箇所に付与することにより，従来の IE に用いられるアノテーションつきコーパスを疑似的に作成する（擬似コーパス）．

2）情報抽出モデルの学習・予測

この擬似コーパスを用いて IE モデルを訓練する．IE モデルには Cheng らが提案した言及箇所抽出とモダリティ分類，関係抽出を同時学習する BERT に基づくモデルを用いる[2]．この IE モデルに症例報告を入力し，出力される関係3つ組を一定のルールに基づいて組み上げることでグラフ構造要約を作成する．

4. 結果

約 15,100 件の症例報告から作成した擬似情報抽出ラベルを訓練:開発:テスト＝14,400:200:500 に分け，IE モデルを訓練，テストした．精度はモデルが出力した関係 3 つ組を，ゴールドのグラフ構造要約を分解して得られたものと比較して算出した(表 1)．

表 1 IE モデルの関係分類の精度

適合率	再現率	F 値
80.5	63.5	69.8

5. 考察

表 2 図 1 の例に対するモデルの予測

全身性エリテマトーデス
- 発熱
- 体重減少
- 関節痛
- 全身倦怠感
- リンパ球/低値
- 関節炎
- 抗核抗体/陽性
- 抗 DNA 抗体/陽性
- 病理組織＝虚血性腸炎
  - 大腸内視鏡＝多発性＊潰瘍@直腸
  - シクロフォスファミド点滴静注療法/有効

混合性結合組織病
- 関節痛
- 抗 RNP 抗体/陽性

表 2 に図 1 に示した症例に対するモデルの予測を示す．混合性結合組織病（赤字)をルートとして予測しており，これは誤りである．一方、青字はゴールドとは違うが，間違いとは言い切れない箇所である．例えば、本症例の関節痛は関節炎の子の可能性が高いが、全身性エリテマトーデスの子でもよく、関節炎と両方の子とするのが妥当だろう．このように、直接の子供ではない子孫との間に親子関係を予測した場合は正解と判定する．しかし、その他の間違いとは言い切れないパターンは現在は不正解と扱っている．より良い評価基準を考案することは今後の課題である．

6. 結語

本研究では，グラフ構造要約を IE 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3つ組抽出の結果は F 値で 69.8 である．今後は技術的改善を行うとともに，本研究を J-CaseMap の拡張に活用することを検討する．

参 考 文 献

[1] 永井良三: AI 時代の"臨床医学のまなざし", 医学のあゆみ Vol.274 No.9, pp703-711, 2020.

[2] Cheng, F., Yada, S., Tanaka, R., Aramaki, E., Kurohashi, S.: JaMIE: A Pipeline Japanese Medical Information Extraction System, Proceedings of the 12th Language Resources and Evaluation Conference, 2022.