

症例テキスト間の論理推論における病名知識補完の試み

村上夏輝¹ 石田真捺¹ 谷中瞳² 戸次大介¹

¹お茶の水女子大学 ²東京大学

{murakami.natsuki, ishida.mana, bekki}@is.ocha.ac.jp hyanaka@is.s.u-tokyo.ac.jp

概要

医療分野には電子カルテなどの症例テキストが多く存在しており、これらを用いた自然言語処理の研究が行われている。その一つに、意味解析・論理推論システム `ccg2lambda` を用いた症例テキストの含意関係認識の研究があるが、既存の推論システムでは、入力文に病名などの医療ドメイン特有の言い換え表現が含まれていた場合、含意関係を正しく判定することができなかった。そこで本研究では定理証明に不足する言い換え表現の候補を病名の固有表現抽出器と万病辞書を用いて特定し、病名の同値関係を公理として補完する手法を提案する。評価実験の結果、提案手法では推論テストセット 108 件のうち、73 件について正しく推論を行うことができた。

1 はじめに

医療分野には画像所見や電子カルテといった電子化されたテキストが多く存在しており、これらを用いた自然言語処理の研究は活発に行われつつある [1, 2]。その一方で、否定や量化といった深い意味を考慮した医療言語処理技術は未だ発展途上にある。

症例テキストを有効活用する手法として、症例テキスト間の含意関係認識が考えられ、その取り組みの一つに石田らのシステム [3] がある。石田らのシステムは、組合せ範疇文法 (Combinatory Categorical Grammar, CCG) [4, 5] に基づく頑健かつ高精度な構文解析と、高階論理に基づく自動推論システムを組み合わせた意味解析・論理推論システムである `ccg2lambda` [6] をベースとした推論システムである。従来の `ccg2lambda` では、複合語が一語として扱われてしまうことで、複合語内の意味的な関係を解析できないという問題点があった。この問題点に対して、石田らのシステムでは `ccg2lambda` に複合語解析モジュールを追加することで解決を図った。石田らの複合語解析モジュールは、`ccg2lambda` の構文解析で得られた CCG 構文木から複合語を抽出し、複

合語意味現象タグ分類モデルを用いて複合語内の意味関係を表す意味現象タグを付与する。そのタグを元に複合語内の意味関係を考慮した意味表示を導出し、複合語を含む症例テキスト間の推論を実現した。

しかし、石田らのシステムでは症例テキストによって病名表記の揺れがあった際に推論を行うことができないという問題がある。例えば、「深部静脈血栓症」という病名には、「DVT」や「ホーマンズ徴候」といった複数の表記が存在し、症例テキストによって表記が異なる。前提文が「DVT 発症を認めた」、仮説文が「深部静脈血栓症を発症した。」である場合、経験的には前提文は仮説文を含意するが、この含意関係を示すためには、「DVT は深部静脈血栓症である」という知識を定理証明の際に補完する必要がある。そこで本研究では、病名の固有表現抽出器と万病辞書 [7] を用いて定理証明に不足する言い換え表現の候補を特定し、病名の同値関係を公理として補完する手法を提案する。また、病名の知識補完が必要な推論テストセットを構築し、提案手法の有効性を評価する。

2 提案手法の全体像

提案手法の全体像を図 1 に示す。石田らのシステムは構文解析、複合語解析、意味解析、定理証明から構成される。本研究では、医療ドメインの知識補完を行うモジュールを石田らのシステムに追加する。知識補完モジュールでは、まず入力文に対して、汎用言語モデルに基づく固有表現抽出器を用いて病名の固有表現抽出を行う。抽出された病名に対して、万病辞書内の病名の出現形欄について完全一致検索を行い、病名と出現形が一致した場合、万病辞書内の病名に関する知識を公理として定理証明器 Coq [8] に追加する。追加する公理は形態素間の意味関係によって異なるため、石田らの複合語意味現象タグ分類モデルによって生成される複合語の意味現象タグを確認し、意味現象タグと万病辞書の知識

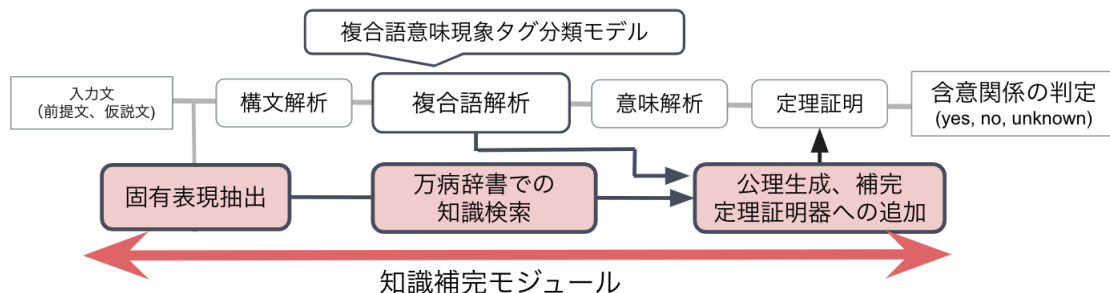


図1 提案手法の全体像

表1 標準病名欄が「深部静脈血栓症」である万病辞書の例

出現形	出現形よみ	ICD10 コード	標準病名	信頼度レベル	頻度レベル
深部静脈血栓症	しんぶじょうみゃくけっせんしょう	I802	深部静脈血栓症	S	85-90%
下肢深部静脈血栓	かししんぶじょうみゃくけっせん	I802	深部静脈血栓症	A	5-10%
DVT	でいぶいてい	I802	深部静脈血栓症	C	90-95%
DVT 徴候	でいぶいていちょうこう	I802	深部静脈血栓症	C	60-65%
ホーマンズ徴候	ほーまんずちょうこう	I802	深部静脈血栓症	C	25-30%

に応じて公理を導出してから補完を行う。

3 病名の知識補完

3.1 万病辞書

本研究における病名の知識補完には万病辞書を用いる。万病辞書は、医療従事者が電子カルテに記録した経過記録と退院サマリから、症状や病名に関連する語を広く抽出したデータである。病名の正式名称だけでなく、略記や英語名を含めた、362,866 件の病名用語が収録されている。

表1に万病辞書の標準病名欄が「深部静脈血栓症」である例を示す。万病辞書は左から、病名の出現形、出現形よみ、ICD10 コード、標準病名、信頼度レベル、頻度レベルが記載されている。ICD10 コードと標準病名は ICD10 対応標準病名マスター¹⁾を参照している。

3.2 病名の固有表現抽出

3.2.1 病名の固有表現抽出データセット

固有表現抽出器の学習を行うために、症例テキストを用いて固有表現抽出データセットを構築した。症例テキストには、J-Stage でオープンアクセス公開されている症例報告論文 PDF から OCR 抽出した症例報告コーパスである J-MedStd-CR²⁾を用いる。本研究では J-MedStd-CR の 2626 文に対して、文中に現れる万病辞書の出現形を人手でアノテーションし

た。以下にアノテーションの例を示す。

```

1 "text" : " 切断前から両側下肢に筋萎縮を
   認めた",
2 "entities": [{
3     "name" : " 筋萎縮",
4     "span" : [10,13],
5     "type" : " 万病辞書"}]

```

3.2.2 固有表現抽出の関連研究

病名の固有表現抽出の関連研究として五井野ら [9] や MedNER[10] がある。五井野らでは、BERT[11] と CRF を用いて病名、症状とそのモダリティ (5 種類) の固有表現抽出とモダリティ推定を行っている。BERT には東北大学が公開している 3 種類の BERT(tohoku-charBERT, tohoku-BERT, tohoku-wwm-BERT) と東京大学が公開している日本語の医療文書で事前学習を行った UTH-BERT³⁾を用いて比較実験を行っている。MedNER では、患者のために書かれた医学研究論文の一種である症例報告 100 文書、放射線科医によって書かれた一種の臨床文書である読影所見 72 文書を学習データとして用いており、少ないリソースを用いた固有表現抽出を行っている。

4 病名の知識補完が必要な推論テストセットの構築

本研究では病名の知識補完が必要なテストセットを構築し、提案手法の評価を行う。表2に推論テストセットの例を示す。テストセットの構築には、症例報告コーパスである J-MedStd-CR 内の病名が現れ

1) <http://www2.medis.or.jp/stdcd/byomei/index.html>

2) <https://sociocom.naist.jp/medtxt/cr/>

3) <https://ai-health.m.u-tokyo.ac.jp/home/research/uth-bert>

表2 推論テストセットの例 (太字箇所は含意関係の判定に必要な病名知識)

前提文	仮説文
CMV腸炎が鑑別診断に上げられた。	⇒ サイトメガロウイルス腸炎が鑑別診断に上げられた。
DLEを合併した。	⇒ 円板状エリテマトーデスを合併した。
患者はホーマンズ徴候陽性であった。	⇒ 患者は深部静脈血栓症陽性であった。

表3 実験に使用する RoBERTa モデル

モデル	事前学習コーパス	Tokenizer
Japanese RoBERTa base(理研)	Wikipedia	MeCab + BPE
japanese-roberta-base(rinna)	Wikipedia + CC-100	Juman++ + sentencepiece
roberta-large-japanese(早稲田)	Wikipedia + CC-100	Juman++ sentencepiece

る文のうち、万病辞書の病名の出現形と標準病名が異なる文を用いる。簡略化した文を仮説文、仮説文の病名箇所を標準病名に変えたものを前提文とした文ペアを人手で作成し、合計 108 ペアからなる推論テストセットを構築した。

5 実験

5.1 実験概要

固有表現抽出の実験概要 固有表現抽出器を作成するにあたって、表 3 に示す 3 つの RoBERTa[12] を用いて固有表現抽出器を構築し、実験を行った。理化学研究所から公開されている Japanese RoBERTa base⁴⁾(以下、理研モデル) は日本語 Wikipedia のみを事前学習に使用しており、Tokenizer には単語分割に MeCab[13]、サブワード分割に BPE[14] を用いている。rinna 社が公開している japanese-roberta-base[15](以下、rinna モデル) と早稲田大学が公開している roberta-large-japanese⁵⁾(以下、早稲田モデル) は日本語 Wikipedia と CC-100 の日本語部分を事前学習に使用している。Tokenizer は、単語分割には Juman++[16]、サブワード分割には sentencepiece[17] を使用している。

固有表現抽出器は、症例テキスト内の 2303 文を学習データとして用い、そのうち 85%を開発データ、15%を検証データとして汎用言語モデルのファインチューニングを行い構築した。この学習データには病名の出現形が 2551 件出現している。評価データには、同じく J-MedStd-CR の症例テキスト内からランダムで抽出した 326 文を用いた。この 326 文は学習データとの重複はない。

推論システムの実験概要 推論システムには石田らのシステムを用いる。石田らのシステムの複合語意味現象タグ分類モデルには、BiLSTM[18] と BERT[11] を用いている。BERT には日本語の Wikipedia のデータを用いて学習された、東北大学が公開しているモデル⁶⁾を使用している。Tokenizer は MeCab+WordPiece である。本研究の提案システムの有効性を評価するために、石田らのシステム+知識補完モジュールありと石田らのシステム+知識補完モジュールなしでの比較を行う。

5.2 病名の固有表現抽出モデルの評価

実験では表 3 の 3 つの BERT に対して、固有表現抽出データセットで学習を行い、評価を行った。また、関連研究の MedNER に対して評価データセットを用いて比較を行った結果を表 4 に示す。病名の予測精度としては理研モデルのスコアが一番高かった。rinna モデルに関しては病名は抽出できていないものの前後の関係ないものまで抽出していた。また、早稲田モデルは抽出された病名が途中で区切られているものが見られた。MedNER に関しては、「壊死を伴う肉芽種」の「壊死」と「肉芽種」や、「疼痛と搔痒」の「疼痛」と「搔痒」といった近くに並んだ病名をまとめて抽出する傾向があった。本研究で作成した評価データセットは、病名单体をつずつアノテーションしたため f1-score が下がったと考えられる。

この実験を経て、提案手法に組み合わせる固有表現抽出器には最も性能の良かった理研モデルを採用する。

4) <https://huggingface.co/liat-nakayama/japanese-roberta-base-20220905>

5) <https://huggingface.co/nlp-waseda/roberta-large-japanese>

6) <https://huggingface.co/cl-tohoku/bert-base-japanese>

表4 病名の固有表現抽出の実験結果

モデル	Japanese RoBERTa base(理研)	japanese-roberta-base(rinna)	roberta-large-japanese(Waseda)	MedNER
accuracy	97.2%	96.4%	97.0%	94.3%
病名の f1-score	81.0%	77.3%	74.6%	66.1%

表5 エラーのタイプと文ペアの例

エラーのタイプ	件数	例
形態素解析によるエラー	13	熱感が出現した。⇒ ほてりが出現した。
構文解析によるエラー	13	四肢疼痛が増悪した。⇒ 四肢痛が増悪した。
固有表現抽出によるエラー	11	interface hepatitisが見られた。⇒ 肝炎が見られた。

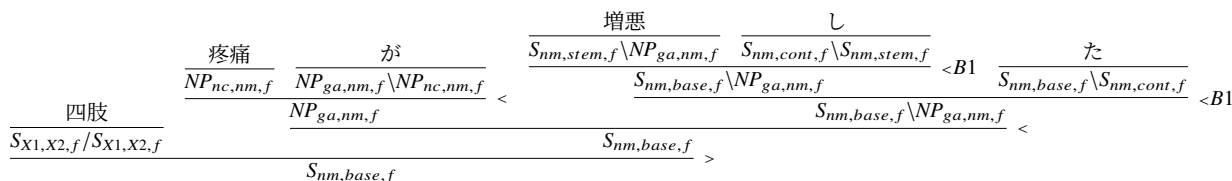


図2 「四肢疼痛が増悪した」の構文解析結果

5.3 知識補完した推論の評価

構築した推論テストセットを用いて、石田らのシステムと、石田らのシステムに知識補完モジュールを追加した提案システムとで推論の精度比較を行った。表6に推論の評価結果を示す。

表6 推論の評価結果

推論システム	精度
石田らのシステム	0/108(0.0%)
石田らのシステム + 知識補完モジュール	73/108 (67.0%)

石田らのシステムでは正しく推論を行える例がなかったのに対して、提案システムでは108件中73件のテストセットについて正しく推論を行うことができた。

提案システムが推論に失敗してしまった例についてエラー分析を行った。表5にエラーのタイプと文ペアの例を示す。形態素解析によるエラーでは、名詞として扱いたい「ほてり」や「すくみ足」の「すくみ」という形態素に対して、形態素解析器であるJanomeは動詞と判定した。そのため、意味解析の際に「ほてる」、「すくむ」といった動詞で扱われてしまい、定理証明で追加した公理が適用されず推論に失敗してしまった。構文解析によるエラーについては、例を図4に示す。「四肢疼痛」は病名であるため「四肢」と「疼痛」は先に結合する必要があるが、構文解析の結果、「疼痛が増悪した」が先に結合し、その後「四肢」と結合している。このように病名部分のCCG木が正しく構築されておらず、正しく推論が行えない例がみられた。固有表現抽出によるエラーでは、「interface hepatitis

is]から「interface」のように、語の途中で途切れて抽出してしまうか、もしくは抽出ができなかったため、正しく推論が行えなかった。

6 おわりに

本研究では、石田らで提案された意味解析・論理推論システムに対して、医療ドメインの知識補完モジュールを追加し、推論システムを構築した。具体的には、病名の固有表現抽出を行ったのち、抽出された病名に対して万病辞書を用いて知識検索を行い、検索結果に基づいた公理を石田らのシステムの定理証明器に追加するという知識補完モジュールを作成した。また、固有表現抽出を行うモデルの学習のために病名の固有表現抽出データセット、医療ドメインの知識が補完された推論システムの評価を行うための推論テストセットを構築した。構築したテストセットを用いて推論システムの評価を行った結果、推論テストセット108件のうち、73件について推論を正しく行うことができた。今後の課題として、固有表現抽出データセット、推論テストセットの拡張を行うとともに、知識補完モジュールの改良によって推論システムのさらなる性能向上を目指す。

謝辞. 本研究の一部は、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）21AC5001、および JST さきがけ JPMJPR21C8、JSPS 科研費 JP20K19868 の支援を受けたものである。

参考文献

- [1] 荒牧英治, 岩尾友秀, 若宮翔子, 伊藤薫, 矢野憲, 大江和彦. 症例検索システムの試行運用に基づいた利用

- 状況に関する基礎的検討. 医療情報学, Vol. 38, No. 4, pp. 245–256, 2018.
- [2] 土井俊祐, 木村隆, 関根正樹, 鈴木隆弘, 高林克日己, 田村俊世. 学会ホームページにおける類似症例検索システムの実運用と評価. 生体医工学, Vol. 49, No. 6, pp. 870–876, 2011.
- [3] 石田真捺, 谷中瞳, 戸次大介. 日本語症例テキストの複合語解析と論理推論. 第36回人工知能学会全国大会論文集. 一般社団法人人工知能学会, 2022.
- [4] Mark J. Steedman. **The Syntactic Process**. The MIT Press, 2000.
- [5] 戸次大介. 日本語文法の形式理論：活用体系・統語構造・意味合成. くろしお出版, 2010.
- [6] Pascual Martínez-Gómez, Koji Mineshima, Yusuke Miyao, and Daisuke Bekki. ccg2lambda: A Compositional Semantics System. In **Proceedings of ACL 2016 System Demonstrations**, pp. 85–90, 2016.
- [7] 荒牧英治, 若宮翔子, 河添悦昌. カルテ情報の自動構造化システムと疾患数理モデルの逐次的構築, 及び, 自動構造化機能を有した入力機構の開発. 医療情報学, Vol. 40, No. 1, pp. 12–13, 2020.
- [8] Yves Bertot and Pierre Castéran. **Interactive Theorem Proving and Program Development: Coq'Art: The Calculus of Inductive Constructions**. Springer Science & Business Media, 2013.
- [9] 五井野琢也, 濱上知樹. Bert を用いた医療文書からの固有表現抽出. 第48回知能システムシンポジウム, 2021.
- [10] Tomohiro Nishiyama, Mihiro Nishidani, Aki Ando, Shuntaro Yada, Shoko Wakamiya, and Eiji Aramaki. NAISTSOC at the NTCIR-16 Real-MedNLP Task. In **Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies**, 2022.
- [11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. **arXiv preprint arXiv:1810.04805**, 2018.
- [12] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. **ArXiv**, Vol. abs/1907.11692, , 2019.
- [13] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to japanese morphological analysis. In **Proceedings of the 2004 conference on empirical methods in natural language processing**, pp. 230–237, 2004.
- [14] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural Machine Translation of Rare Words with Subword Units. In **Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 1715–1725. Association for Computational Linguistics, 2016.
- [15] 趙天雨, 沢田慶. 日本語自然言語処理における事前学習モデルの公開. 人工知能学会研究会資料 言語・音声理解と対話処理研究会, Vol. 93, pp. 169–170, 2021.
- [16] Arseny Tolmachev, Daisuke Kawahara, and Sadao Kurohashi. Juman++: A morphological analysis toolkit for scriptio continua. In **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations**, pp. 54–59. Association for Computational Linguistics, 2018.
- [17] Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. **arXiv preprint arXiv:1808.06226**, 2018.
- [18] Peilu Wang, Yao Qian, Frank K Soong, Lei He, and Hai Zhao. Part-of-speech tagging with bidirectional long short-term memory recurrent neural network. **arXiv preprint arXiv:1510.06168**, 2015.