The 36th Annual Conference of the Japanese Society for Artificial Intelligence, 2022

診療テキストの構造化に向けた症例報告コーパスからの情報抽出

Information Extraction from Japanese Case Report Corpus for Structuring Clinical Texts

柴田 大作[*1] 河添 悦昌[*1] 篠原 恵美子[*1] 嶋本 公徳[*1]
Daisaku Shibata Yoshimasa Kawazoe Emiko Shinohara Kiminori Shimamoto

[*1] 東京大学大学院 医学系研究科
Graduate School of Medicine, The University of Tokyo

[Background] Significant information related to symptoms and findings of the patients is often written in a free-text form in clinical texts. To utilize these texts, information extraction using Natural Language Processing is required. [Objective] In this study, we evaluated named entity recognition (NER) and relation extraction (RE) performances with machine learning methods. We utilized the Japanese Case report corpus, which has manually annotated 70 type of entities and 35 type of relations. [Method] This study utilized the aforementioned corpus containing 183 cases. Having pre-processed them, we finally used 182 cases consisting of 2,172 sentences. Furthermore, a machine learning model based on Bidirectional Encoder Representations from Transformers was used. [Result] The results revealed that the maximum micro-averaged F1 scores of NER and RE were 0.931 and 0.826, respectively. [Discussion] We obtained comparable results to previous studies. Hence, these results could be substantial accuracies as baselines.

1. はじめに

実臨床の場において、希少・難治性疾患は種類が多く頻度が低いことから、医師にとって未経験の疾患が多く存在し、疾患の見落としが生じる可能性がある。そこで、蓄積された電子カルテデータから情報検索技術を活用し、疾患のスクリーニングを行い、類似した症例を抽出することで疾患の見落としを防止できる可能性がある。電子カルテデータに記録される情報は、1) 年齢、性別、血液型、アレルゲン物質などの基本情報、2) 喫煙や飲酒などの生活習慣、3) 家族歴、4) 過去や現在の病名、5) 血液や尿などの検体検査結果、6) 病理検査や放射線検査の検査所見などが含まれる。これらの情報は構造化データとして記録されるものもあるが、放射線レポートや退院サマリなどの診療テキストにフリーテキストとして記載される場合も多い。疾患のスクリーニングにおいては、これらの情報全てを構造化データとして保持することが有益であるが、フリーテキストから必要な情報を手動で抽出することは時間的・金銭的なコストが大きい。そのため、自然言語処理技術を用いた診療テキストからの情報抽出が必要となる。

診療テキストからの情報抽出に関する研究はいくつか報告されており、[Isar 19]では、退院サマリに出現する病名や症状、医薬品などの表現に対してアノテーションが付与された i2b2 2010 データセット [Uzuner 11] などから Bidirectional Encoder Representations from Transformers（BERT）[Devlin 18]を用いた固有表現抽出を行うことで、従来手法よりも高い精度が得られることを報告している。また、[Mulyar 21]では、退院サマリに出現する医薬品と薬物有害事象についてアノテーションした n2c2 2018 データセットや退院サマリに出現する患者名、医師名や病院名などの個人情報に対してアノテーションが付与された i2b2 2014 データセット [Amber 15]などから、固有表現抽出などのタスクを BERT により実施した結果を報告している。

英語では診療テキストにアノテーションが付与されたデータセットがいくつか利用可能であることから、さまざまな分野の研究者が機械学習を用いた手法を考案し、その精度を報告している。これはコミュニティの発展に大きく寄与するものであるが、日本語では利用可能なデータが少ないため、そのような活動は少なく、コミュニティの発展を妨げる一因となっている。

そのため本研究では、2021 年に公開された日本語の症例報告コーパスを用いた機械学習による固有表現抽出と関係抽出を行い、本コーパスにおける現状のベースラインとなる精度を報告し、得られた精度の妥当性や機械学習モデルの問題点について考察する。

連絡先：〒113−8655 東京都文京区本郷 7−3−1 中央診療棟 2 22 世紀医療センター8 階 医療 AI 開発学講座

2. 方法

2.1 実験材料

実験材料として、日本語の症例報告コーパス（以後、iCorpus）を用いる [篠原 21]。iCorpus は、厚生労働省の指定難病名と「例」という文字列をタイトルに含み、2000 年以降に出版された症例 183 件のテキストから構成されている。また、70 種類の固有表現と 35 種類の関係が文字単位でアノテーションされており、各文書は改行で文単位に分割されている。iCorpus のアノテーション例を図 1 に、統計情報を表 1 に、重要な固有表現と関係の詳細を表 2 に示す。なお、コーパスは東京大学大学院医学系研究科の医療 AI 開発学講座の Web ページにて公開されており、本研究では iCorpus_202111 を用いる。

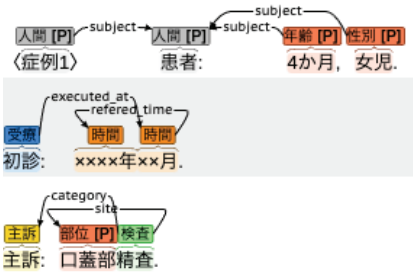

図 1：iCorpus のアノテーション例

表 1：iCorpus の統計情報

文書数		183
固有表現の種類		70
関係の種類		35
1 文	文字数（S.D）	1,915（696）
	単語数（S.D）	972（330）

書毎	固有表現数（S.D）	394（129）
	関係数（S.D）	387（127）

表 2:iCorpus にアノテーションされている固有表現と関係（一部）

固有表現	説明	例
state	患者の状態全般を示す表現	吐き気、萎縮症、糖尿病
body	特定の人体部位を示す表現	肝、手足、眼瞼結膜
item	患者の状態を表すために参照される項目	血糖値、HbA1c、食欲
PN_Positive	患者の状態があることが明示される表現	気付き、認め、示し
value	検査値など身体や検体を測定し得られる数値	7.5、20、1
unit	数値との組で表される単位	%、mg/日、kg
time	時間軸の特定の場所に位置する時点や時区間	直後、元来、7年前
関係	説明	例
value_of	source が target の値である	**身長** (target)は **170** (source)cm
site	source が target の部位である	**四肢** (target)の**筋力** (source)低下
unit	source は target の単位である	身長 **170** (target)**cm** (source)であり
method	source が target (の方法)により得られる	**聴診** (target)上、**異常** (source)なし
executed	target の実施が(source)によって明示される.	**経口投与**(target)が**開始された**(source)
reason	source と判断された根拠は target である	**蜂窩織炎**(target)から**菌血症**(source)を疑われ

2.2 実験

(1) コーパスの前処理

実験を行うために必要な前処理を iCorpus に対して実施する。日本語のテキストは単語境界が明確でないため、事前に形態素解析器を用いて単語単位に分割する必要がある。形態素解析は辞書に登録された単語に基づいて行われるが、この際、図 2 の(a)と(b)に示すように固有表現のアノテーションスパンと形態素解析により得られた単語の間にギャップが生じ、固有表現抽出のためのラベリングが困難になる場合がある。この問題は直ちに解決することが難しいため、本研究では事前にテキストを固有表現単位に分割し、その後、各固有表現に対して形態素解析を実施する (図 2 の(c))。これは未知のテキストに対する処理とは乖離していることに留意が必要である。

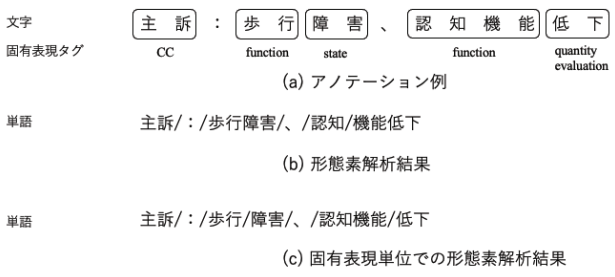

図 2:コーパスの前処理の例

(2) タスク設定

機械学習による固有表現抽出（Named Entity Recognition: NER）と関係抽出（Relation Extraction: RE）の精度評価を行う。文の各単語に最適な固有表現タグを付与するタスクを NER、単語間の関係ラベルを分類するタスクを RE とし、NER には BERT の最終層に Conditional Random Fields（CRF）を接続した BERT-CRF を、RE には Ma らによって提案された Self-Attention ベースの手法 [Ma 20]を使用する。また、NER と RE を同時に実施する Joint 型と NER の後に RE を行う Pipeline 型の 2 種類のモデルを用いる。Joint 型のモデルの詳細を図 3 に示す。

(3) 学習と予測

Joint 型のモデルの学習では、NER により算出される損失と RE により算出される損失の平均値を同時に最小化し、Pipeline 型のモデルではそれぞれを最小化するようパラメーターを学習した。両モデルにおいて、固有表現タグの埋め込み表現は、訓練時は正解の固有表現タグ、予測時はモデルが予測した固有表現タグから取得した。

(4) 実験設定

BERT へ入力する単語数には上限があるため、本研究では入力文の単語数を最大で 510 単語とし、これを超える単語数から構成される文は実験データから除外した。そのため、最終的な実験材料として 182 症例、2,171 文を使用した。

NER は固有表現の最初の単語を Begin、固有表現の最初以外の単語を Inside、固有表現以外の単語を Outside とラベリングする IOB2 形式によりラベリングを行なった。実験では、日本語の診療テキストで事前学習された UTH-BERT [Kawazoe 21]を使用し、最適化関数に AdamW を使用し、バッチサイズ は 1、BERT の学習率は 3e-5、それ以外は 1e-3 に設定し、上記以外のパラメーターは全てデフォルト値を使用した。

(5) 評価方法

評価指標として、5 分割交差検証による Macro-F1、Micro-F1 を算出した。交差検証において、訓練データ、検証データとテストデータの分割は症例単位で実施し、訓練データに含まれる症例のうち 20%を検証データとして使用した。NER と RE の固有表現タグと関係ラベルごとの評価方法をそれぞれ以下に示す。

2.2.5.1. 固有表現抽出の評価

NER では CoNLL-2000 [Sang 00]で使用されている評価方法を使用した。固有表現タグごとの F 値は Precision と Recall の調和平均で算出し、Precision はモデルが固有表現と予測した表現の中で，実際に固有表現だった表現の割合で、Recall は固有表現である表現の中で，固有表現と予測された表現の割合で算出される。

2.2.5.2. 関係抽出の評価

RE ではある文において、ある単語 w_i から単語 w_j への正解の関係ラベルが *rel* であったときに（ある単語が 2 つ以上の単語から構成される固有表現である場合は先頭の単語のみ考え、関係がない場合は *None* という関係ラベルを持つ）、モデルの予測も *rel* であった場合は True Positive（TP）とした。また、本来 *rel* であるものを誤って *rel* 以外と予測した場合は False Negative（FN）とし、本来 *rel* でないものを誤って *rel* と予測した場合が False Positive（FP)とし、Precision と Recall を算出し、両者の調和平均より F1 を算出した。

なお本実験では、NER の予測結果が間違っていた場合でも、RE の予測結果が正しければ RE は正解とした。

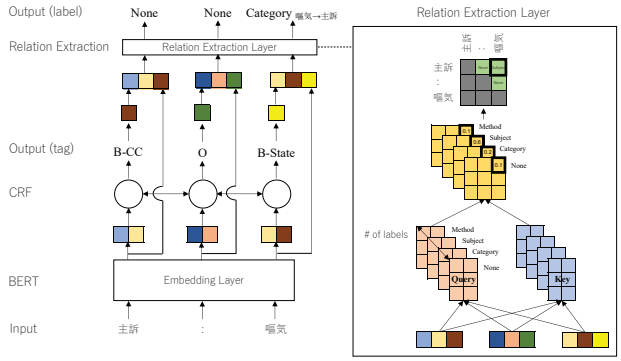

図 3：Joint 型の機械学習モデルの例（入力文を「主訴:嘔気」とし、関係ラベルは *None*、*Subject*、*Category*、*Method* の 4 種類のみとした場合の例を示す）

3. 結果

実験結果を表 3 に示す。表 3 より NER と RE の両タスクにおいて、Micro-F1 と Macro-F1 共に Pipeline 型のモデルの精度が高いことが確認された。

表 3 実験結果

	NER		RE	
	Joint	Pipeline	Joint	Pipeline
Micro-F1	0.926	**0.931**	0.825	**0.826**
Macro-F1	0.753	**0.779**	0.640	**0.643**

4. 考察

4.1 固有表現抽出と関係抽出の精度に関する考察

表 3 より、NER の Micro-F1 は最大 0.931、Macro-F1 は 0.779 であり、RE の Micro-F1 は 0.826、Macro-F1 は 0.643 であることが確認された。先行研究において頻繁に用いられる i2b2 2010 データセットにおける現状の最高精度は exact matching F1 で 90.25 [Si 19]である。また ADE コーパス [Gurulingappa 12]は medical case reports に出現する 2 種類の固有表現と1種類の関係についてアノテーションされたものであり、現状の最高精度は NER が Macro-F1 で 91.3 [Yan 21]、RE が 83.74 [Crone 20]である。そして、コーパスは公開されていないが、Cheng らは 9 種類の固有表現と 10 種類の関係がアノテーションされた日本語の読影レポートと診療録から情報抽出を行い、NER は Micro-F1 でそれぞれ 95.65/85.49、RE も同様に 86.53/71.04 の 精度で抽出できることを報告 [Cheng 21]している。

本研究で使用した iCorpus は、先行研究で使用されているコーパスと比較した場合、アノテーションの粒度が細かく、固有表現タグと関係ラベルの数も多いため、情報抽出のタスクとしてはより難しいものであると考えられる。しかし一方で、先行研究の精度と比較した場合、固有表現タグや関係ラベルの数が大きく異なるにも関わらず、先行研究と遜色ない精度を達成できていることから、ベースラインとなる精度としては十分なものであると考えられる。

4.2 誤り分析

コーパスにおいて出現頻度の高い固有表現タグと関係ラベルの一部を選択し、誤り分析の対象とした。対象とした固有表現タグは *ent: state*、*body*、*item* 関係ラベルは *rel: value_of*、*site*、*unit*、*method* である。

(1) 固有表現抽出

NER の実験結果から正解タグを基準に作成した混合行列を図 4 に示す。なお、固有表現タグは 4.2 節で述べた 3 種類に限定し、固有表現タグの抽出範囲を誤った事例や固有表現ではないと誤って予測した事例などはまとめて *Outside* とした。また、限定した 3 種類の固有表現タグ以外のタグと誤って予測した事例 (例えば *ent: body* を *ent: activity* と予測)は扱わない。なお本節では、誌面の都合上、*ent: state*、*body* のみについて述べる。

ent: state では *Outside* と予測した事例が最も多く 681 件あった。これには「小脳/##皮質/萎縮症」の「小脳/##皮質」を *ent: body*、「萎縮症」を *ent: state* と 1 つの固有表現を 2 つの異なる固有表現とした事例や 2 つの異なる固有表現である「正常/出産」(それぞれ *ent: state*)をまとめて *ent: state* と予測した事例などがあった。次いで、*ent: item* と誤って予測した事例が 72 件であり、「内容物」や「N/AF/LD」などの固有表現に対して誤って予測した事例が確認された。そして、*ent: body* と誤って予測した事例が 36 件であり、「輪状」や「気管/気管支」(それぞれの単語を *ent: body* と予測)などの固有表現に対して誤って予測している事例が確認された。

ent: body でも同様に *Outside* と予測した事例が最も多く 290 件あった。例えば、それぞれ *ent: body* をもつ固有表現である「下肢/静脈」をまとめて *ent: body* をもつ 1 つの固有表現として予測した事例や、反対に *ent: body* をもつ 1 つの固有表現である「視床/枕」をそれぞれ異なる固有表現として *ent: body* を予測した事例があった。次いで、*ent: state* と誤って予測した事例が 32 件あり、「膜/様/部」や「上大静脈」などの固有表現に対して誤って予測した事例が確認された。そして、*ent: item* と誤って予測した事例が 7 件あり、「背景/肝」や「上腕二頭筋」などの固有表現に対して誤って予測した事例が確認された。

これらの予測を誤った事例には、学習データ中の出現頻度が著しく低い固有表現である場合や文脈によって固有表現タグが変化する場合、またより小さいもしくは大きいスパンの固有表現が存在する場合があり学習が困難であると考えられる。そのため今後は、データ拡張などを用いた学習データの増加やよりタスクに適した事前学習済みの言語モデルを適用していくことなどが必要であると考えられる。

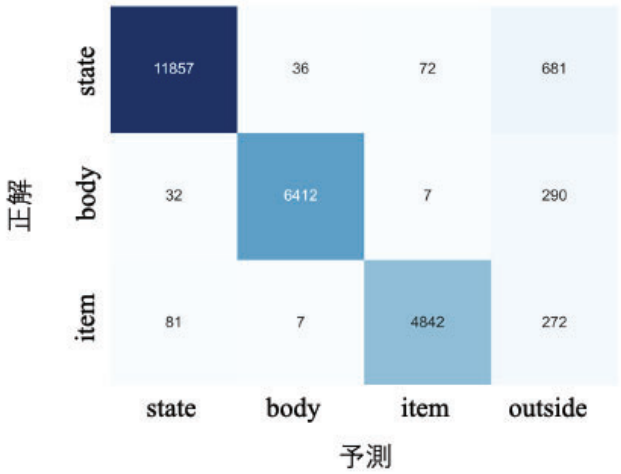

図 4：NER の混合行列

(2) 関係抽出

RE の実験結果から正解タグを基準に作成した混合行列を図 5 に示す。なお、関係ラベルは 4.2 節で述べた 4 種類に限定した。また、限定した4種類の関係ラベル以外のラベルと誤って予測した事例 (例えば、*rel: site* を *rel: executed_at* と予測)は扱わ

ない。なお本節では、誌面の都合上、*rel: value_of*、*site* のみについて述べる。

rel: value_of では、本来 *rel: value_of* の関係を持つ固有表現間に誤って *rel: None* と予測した事例が最も多く 2,752 件あった。出現頻度が多かった事例として、「口蓋/裂」は「裂」から「口蓋」に *rel: value_of* を持つが、これを *rel: None* と予測した事例、また「非/月経」は「非」から「月経」に *rel: value_of* を持つが、これを *rel: None* と予測した事例などがあった。次いで *rel: site* と誤って予測した事例が多く、183 件あった。出現頻度が高かった事例として、「脊椎/側弯」は「側弯」から「脊髄」に *rel: value_of* を持つが、これを *rel: site* と予測した事例、また「大動脈/解離」は「解離」から「大動脈」に *rel: value_of* を持つが、これを *rel: site* と予測した事例などがあった。

rel: site では、本来 *rel: site* の関係を持つ固有表現間に誤って *rel: None* と予測した事例が最も多く 962 件あった。出現頻度が多かった事例として、「腹部」と「圧痛」は「圧痛」から「腹部」に *rel: site* を持つが、これを *rel: None* と予測した事例、また「正中神経」と「低下」は「低下」から「正中神経」に *rel: value_of* を持つが、これを rel: None と予測した事例などがあった。次いで *rel: value_of* と誤って予測した事例が多く、140 件あった。出現頻度が高かった事例として、「憩室/穿通」は「憩室」から「穿通」に *rel: site* を持つが、これを *rel: value_of* と予測した事例、また「びらん」と「接触痛」は「接触痛」から「びらん」に *rel: site* を持つが、これを *rel: value_of* と予測した事例などがあった。

これらの予測を誤った事例には、NER の予測を誤ったため、関係ラベルの予測を誤ったと考えられる事例がいくつか確認された。そのため、RE の精度向上には NER の精度向上も同時に必要であると考えられる。

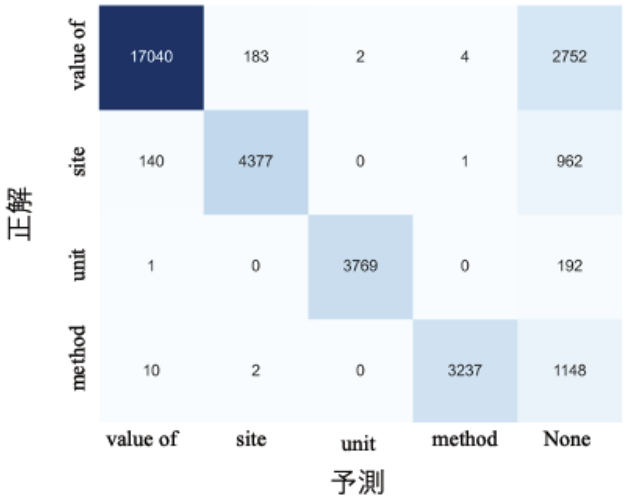

図 5:RE の混合行列

5. まとめ

本研究では症例報告に出現する 70 種類の固有表現と 35 種類の関係がアノテーションされた症例報告コーパスを用いた機械学習による固有表現抽出と関係抽出を行なった。その結果、固有表現抽出は Micro-F1 で最大 0.913、関係抽出は 0.826 の精度が得られ、ベースラインとなる精度を提示した。

今後は本研究により得られた機械学習モデルを用い、実際の医療現場で作成されるテキスト(例:退院サマリ)に対して精度評価を行うことで本コーパスとモデルの有用性について調査する予定である。

参考文献

[Isar 19] Nejadgholi, Isar, et al.: Recognizing umls semantic types with deep learning, Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019). 2019.

[Uzuner 11] Uzuner, Özlem, et al.: 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. Journal of the American Medical Informatics Association 18.5, pp. 552-556 (2011).

[Devlin 18] Devlin, Jacob, et al.: Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805, 2018.

[Mulyar 21] Mulyar, Andriy, Ozlem Uzuner, and Bridget McInnes: MT-clinical BERT: scaling clinical information extraction with multitask learning, Journal of the American Medical Informatics Association 28.10, pp. 2108-2115 (2021).

[Amber 15] Stubbs, Amber, Christopher Kotfila, and Özlem Uzuner.: Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task Track 1, Journal of biomedical informatics 58, pp. S11-S19 (2015).

[篠原 21] 篠原 恵美子, et al.: 医療テキストに対する網羅的な所見アノテーションのためのアノテーション基準の構築. 第 25 回日本医療情報学春季学術大会, 2021.

[Ma 20] Youmi Ma, et al.: Named Entity Recognition and Relation Extraction using Enhanced Table Filling by Contextualized Representations, arXiv preprint arXiv: 2010.07522, 2020.

[Kawazoe 21] Kawazoe, Yoshimasa, et al.: A clinical specific BERT developed using a huge Japanese clinical text corpus, Plos one 16.11, (2021).

[Sang 00] Sang, Erik F., and Sabine Buchholz.: Introduction to the CoNLL-2000 shared task: Chunking. arXiv preprint cs/0009008 (2000).

[Si 19] Si, Yuqi, et al.: Enhancing clinical concept extraction with contextual embeddings, Journal of the American Medical Informatics Association 26.11, pp. 1297-1304, (2019).

[Gurulingappa 12] Gurulingappa, Harsha, et al.: Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports, Journal of biomedical informatics 45.5, pp. 885-892, (2012).

[Yan 21] Yan, Zhiheng, et al.: A Partition Filter Network for Joint Entity and Relation Extraction, Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 185-197, 2021.

[Crone 20] Crone, Phil.: Deeper Task-Specificity Improves Joint Entity and Relation Extraction, arXiv preprint arXiv:2002.06424, 2020.

[Cheng 21] Cheng, Fei, et al.: JaMIE: A Pipeline Japanese Medical Information Extraction System, arXiv preprint arXiv:2111.04261, 2021.

[Henry 21] Henry, Sam, et al.: 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records, Journal of the American Medical Informatics Association 27.1, pp. 3-12 (2020).