Short Report

# 深層学習による医療文書からの病名と医療行為の抽出

堂坂浩二，石井雅樹，伊東嗣功

*秋田県立大学システム科学技術学部情報工学科*

医療分野において，電子カルテは様々な医療文書から成り，病院の業務改善，医療の質の向上につながる情報源である．電子カルテの医療文書を解析することにより，電子カルテ監査といった医療業務の生産性向上や，薬の副作用に関する知識の獲得などに役立てることができる．本研究では，医療文書から病名と医療行為を抽出する固有表現抽出器を開発し，評価を行った．従来の日本語の医療言語処理では，病名のみをアノテートしたコーパスの開発が進められ，そのコーパスに基づいて病名の固有表現抽出器が開発されてきた．しかし，電子カルテ監査等の業務に適用するためには，病名だけでなく，手術，処置，検査，投薬等の医療行為を認識することが重要である．そこで，本研究では，小規模ではあるが，病名だけでなく医療行為をアノテートしたコーパスを作成し，病名と医療行為の固有表現抽出器を深層学習により構築した．病名と比べると医療行為の抽出性能は低いものの，アノテーション付きコーパスを整備することにより，深層学習を使って医療文書から病名と同様に医療行為を抽出できる可能が示された．

**キーワード：**医療言語処理，電子カルテ，医療概念抽出，固有表現抽出

医療分野において，電子カルテは，医師による診療録・処置記録・退院サマリ，看護師による看護記録などの医療文書から構成され，医療の質の向上，病院の業務改善，薬の副作用の分析，医療従事者と患者間での診療情報の共有等につながる重要な情報源である（高間ら, 2015; 山田, 2017）．大学病院規模の医療機関では電子カルテ文書が月に 20 万件以上もやりとりされると言われる（荒牧ら, 2017）．人口減少社会において，膨大な電子カルテ文章を活用することで，医療業務の生産性向上，医療の質の確保に役立てることは重要な課題である．

医療文書を解析し，情報を抽出する技術は医療言語処理と呼ばれる．医療文書を解析する際，最も基本的な処理は，文書中から医療概念を表す言語表現（固有表現）を抽出し，固有表現を病名，症状，手術，処置，検査，投薬などのカテゴリに分類する処理である．この処理は固有表現抽出と呼ばれる．本研究は医療文書から医療概念を表す固有表現を抽出することを目的とする．

医療概念の固有表現を抽出する手法を開発するためには，医療概念をアノテートした医療文書データ（以後，医療コーパスと呼ぶ）が必要である．海外では， i2b2 NLP Challenge というワークショップが 2006 年から医療コーパスの整備を進めている（Uzuner, 2008）．一方，我が国においては，日本語診療録を処理するための医療文書データの整備は 2012 年頃から始まった．日本語ワークショップ MedNLP（Morita et al., 2013; Aramaki et al., 2014）により，「GSK 2012-D 模擬診療録テキスト・データ」（https://www.gsk.or.jp/catalog/gsk2012-d/）や NTCIR MedNLP-11 コーパス（Aramaki et al., 2014）などのアノテーション付きコーパスが整備された．これらは数十件程度の文書から成る小規模なコーパスであったが，現在 45,000 件の文書から成る大規模な医療コーパスの整備も進められている（荒巻ら, 2018）．しかし，これらの日本語医療コーパスは，医療概念のうち病名や症状（これ以後，病名と呼ぶ）のみをアノテートしたコーパスであり，そのコーパスに基

責任著者連絡先：堂坂浩二　〒015-0055 由利本荘市土谷字海老ノ口 84-4 公立大学法人秋田県立大学システム科学技術学部情報工学科，E-mail: dohsaka@akita-pu.ac.jp

づいて開発された固有表現抽出器も病名とその事実性のみの抽出を行うものである（矢野ら, 2017）.

病名の抽出は最優先で取り組むべき重要なタスクであるが（荒巻ら, 2018），医療事務の生産性向上や薬の副作用の分析等の目的のためには，病名だけでなく，手術，処置，検査，投薬といった医療行為の抽出も不可欠である．例えば，電子カルテ監査業務では，患者の受診から退院までの診療エピソードを記録した診療録，手術・処置記録，診断群分類などの間の整合性を検証する必要がある.多大なマンパワーが必要な業務であり，生産性の向上が望まれている.そのためには,医療文書から病名だけでなく，手術・処置・投薬といった医療行為を抽出し，それらの事象に関して文書間で矛盾がないことを検証しなければならない.

以上の観点から，本研究は，医療文書から病名に加えて医療行為の固有表現を抽出する手法を開発することを目指す．そのためには，病名だけでなく医療行為もアノテートした医療コーパスが必要であるが，そうした日本語医療コーパスは存在しない．そこで，小規模ではあるが，医療行為をアノテートしたコーパスを作成し，病名と医療行為の固有表現抽出器を深層学習により構築し，その評価を行った.

以下において，まず，医療概念の固有表現抽出の方式について説明する．次に，病名と医療行為をアノテートした医療コーパスの作成について述べ，深層学習により医療文書から病名と医療行為を抽出する固有表現抽出器について評価した実験と結果について述べる.

## 医療概念の固有表現抽出

### 固有表現抽出

医療概念には様々なものがあるが，本研究では，次のように病名と医療行為の 2 種類の医療概念に着目する.

- **病名**: 病名や症状
- **医療行為**: 手術，処置，検査，投薬

先に述べたように，電子カルテ監査等の業務に適用するためには，医療文書から病名だけでなく医療行為を抽出することが不可欠である.

2016 年 3 月： S 状結腸癌で S 状結腸切除術
**時間** **病名** **医療行為**
施行．リンパ節, 肝臓, 腹膜に転移あり.
**病名**

*図 1* 病名と医療行為の固有表現の例

| | |
|---|---|
| 2016 年 | B-TIM |
| 3 月 | I-TIM |
| ： | O |
| S | B-BYO |
| 状 | I-BYO |
| 結腸 | I-BYO |
| 癌 | I-BYO |
| で | O |
| S | B-MEA |
| 状 | I-MEA |
| 結腸切除術 | I-MEA |
| 施行 | O |
| 。 | O |
| リンパ節 | O |

*図 2* BIO 形式による固有表現のラベル付与の例

図 1 に病名と医療行為を表す固有表現の例を示す. 図 1 では下線によって固有表現の境界を示し，その下に固有表現のクラスを書いている．ここでは，固有表現クラスとして，病名と医療行為以外に時間のクラスも示している．固有表現抽出とは，このように文書から固有表現の境界を特定し，そのクラスを識別することを言う.

日本語文章が与えられたとき，固有表現抽出を行うためには，まず文章を形態素解析し，語に分解する．どのような語に分解されるかは，形態素解析器に登録されている辞書による．形態素解析によって分解された語の列が入力されると，固有表現抽出器は，語の列に対して固有表現のラベルを付与する．固有表現のラベルの付与の形式として BIO 形式がある．その例を図 2 に示す．図 2 の左の列は形態素解析の結果分解された語の列である．第 2 列が BIO 形式のラベルである．「BYO」，「MEA」，「TIM」はそ

れぞれ病名，医療行為，時間の固有表現クラスを示す．クラスの前の「B-」はそのクラスの固有表現の境界が開始したことをし，「I-」はそのクラスの固有表現の境界が継続していることを示す．「O」は固有表現の境界の外を示す．

### 深層学習による固有表現抽出

固有表現抽出の難しさの一つは，形態素解析による語の分割と固有表現の境界が一致しないことにある．図２では，「S 状結腸癌」という文字列が，形態素解析により「S」，「状」，「結腸癌」という語に分解されている．このように形態素解析器による語の分割を固有表現の境界と一致させることは難しい．また，文書には大量の固有表現が現れ，その表記も揺れるため，すべての固有表現を形態素解析の辞書として登録することは困難である．以上の問題を解決するため，BIO 形式によってラベル付けしたコーパスを用意し，コーパスから機械学習により固有表現抽出器を学習するという方式がとられる．

本研究では，Lample ら（2014）が提案した深層学習による固有表現抽出器を利用する．この方法では，文章を形態素解析によって単語に分割した後，各単語は単語分散表現と呼ばれるベクトルして表現される．図 3 に示すように，単語分散表現は双方向 LSTM（Long Short Term Memory）（Graves et al., 2005）に入力される．双方向 LSTM は文章を前から読んだときの文脈と後ろから読んだときの文脈の双方を考慮して，入力単語に付与すべき固有表現ラベルを予測する．双方向 LSTM の出力は，CRF（条件付確率場）に入力される．CRF は，B-MEA の直後に I-BYO は続かないなど，固有表現ラベル間の依存関係を考慮して，各ラベルの確率を出力する．

この方法では，単語を単語分散表現と呼ばれるベクトルで表現する．まず，固有表現のアノテーション付きコーパスを使って，単語の綴りの情報を考慮した単語分散表現を学習する．綴り（文字の並び）としての病名らしさ，医療行為らしさを学習することにより，未知語に頑健であるという特徴をもつ．さらに，アノテーション無しの大規模コーパスから教師無し学習された訓練済みの単語分散表現を利用することで，固有表現抽出の性能が向上することが

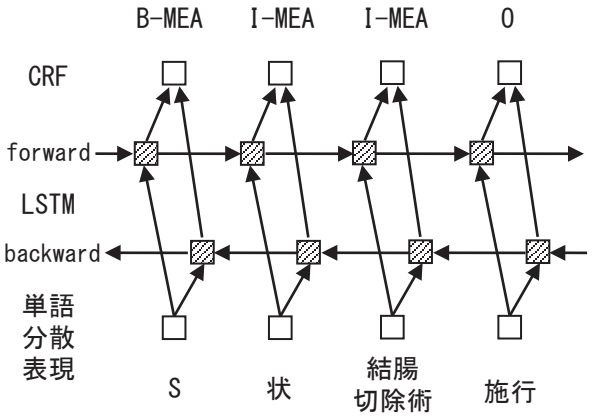


*図3*　双方向 LSTM と CRF による固有表現抽出モデル

示されている．

### 医療コーパスの作成

本研究では，小規模ながら病名だけでなく医療行為をアノテートした医療コーパスを作成した．アノテーションには世界保健機関（WHO）が作成した疾病及び関連保険問題の分類体系である ICD（WHO, 1992）を利用した．

第一に，NTCIR MedNLP コーパス（Aramaki et al., 2014）を利用して，医療行為のアノテーションを付与した．このコーパスは病名と時間表現がアノテーション済みである．このコーパスに対して非医療従事者が ICD-9-CM を参考に医療行為だと判断した言語表現にアノテーションを付与した．このコーパスを「MedNLP 手動コーパス」と呼ぶ．

第二に，診療情報管理士のためのテキストである「ICD コーディングトレーニング第 2 版」の診療録を書き起こし，診療録に対応する文書を抜き出し，アノテーション無しの生コーパスを作成した．次に，日本語病名抽出システム MedEX/J（矢野ら, 2017）を使って，生コーパス解析し，自動的に病名にアノテーションを付与した．さらに，表 1 に示すように，「ICD コーディングトレーニング第 2 版」では，ICD-9-CM に基づいて，診療録にコーディングすべき医療行為名が表形式で記載されている（鳥羽, 2006, pp.293-294）．この医療行為から生成したパターンとのマッチングにより，コーパスに自動的に医療行為

表 1　*ICD コーディングトレーニングの例け*

| | |
|---|---|
| 診療録 | 既往なし。. . . 心エコーで急性突発性心外膜炎の診断. . . |
| 診断名 | 急性突発性心外膜炎　I30.0 |
| 医療行為 | 心エコー　88.72 |

のアノテーションを付与した．時間表現は，手動で作成した正規表現とのパターンマッチングにより自動的にアノテーションを付与した．この結果生成したコーパスを「ICD 自動コーパス」と呼ぶ．

第三に，「ICD 自動コーパス」に対して，非医療従事者が，ICD-10, ICD-9-CM に基づいて，病名と医療行為のアノテーションを見直し修正した．この結果生成されたコーパスを「ICD 手動コーパス」と呼ぶ．

表 2 にそれぞれのコーパスの文の数，病名・医療行為・時間の各固有表現クラスの出現数を示す．「MedNLP 手動コーパス」は医療行為の出現数が少ない．元にしている NTCIR MedNLP コーパス（Aramaki et al., 2014）は退院サマリであるが，手術・処置・検査・投薬等の医療行為に関する言及が少なったためである．

表 2　*作成したコーパス*

| コーパス | 文 | 病名 | 医療行為 | 時間 |
|---|---|---|---|---|
| MedNLP 手動 | 1546 | 1141 | 83 | 245 |
| ICD 自動 | 2223 | 1669 | 430 | 786 |
| ICD 手動 | 2223 | 1742 | 822 | 786 |

## 評価実験

### 実験方法

本研究では, 双方向 LSTM と CFF による固有表現抽出（Lample et al., 2016）の実装として，anaGo（https://github.com/Hironsan/anago）を使った．固有表現のアノテーションが付与されていない大規模コーパスから学習された訓練済み単語分散表現として，日本語版 Wikipedia の本文全文から学習した「日本語 Wikipedia エンティティベクトル」（鈴木ら, 2016）を使用した．次元数は 100 次元のベクトルを使った．

コーパスは，「MedNLP 手動コーパス」，「ICD 自動コーパス」，「ICD 手動コーパス」に加えて，「MedNLP 自動コーパス」と「ICD 手動コーパス」を結合したコーパス，「MedNLP 手動コーパス」と「ICD 手動コーパス」を結合したコーパスの 5 種類に関して評価を行った．コーパスごとに，コーパスを 5 分割し，5 分の 4 を訓練データに，残りの 5 分の 1 の半分をバリデーションデータに，残り半分をテストデータに使い，訓練とテストを行った．バリデーションデータを使った訓練を 15 回繰り返した結果学習されたモデルを使って，テストデータで評価した．これを 5 回繰り返し，固有表現クラス（病名，医療行為，時間）ごとに，固有表現抽出の精度，再現率，F 値のそれぞれのマクロ平均をとった．実行環境は以下の通りである．

- CUDA 9.0
- cuDNN 7.1
- Keras 2.2.4
- 形態素解析: mecab-ipadic-NEologd（佐藤ら, 2017）
- GPU: GeForce GTX TITAN X 11GB

### 評価結果と考察

表 3 に評価実験の結果を示す．各コーパスについて，訓練済み単語分散表現の使用の有無に分けて，固有表現クラスごとの精度（P），再現率（R），F 値（F）を示した．訓練済み単語分散表現の使用有りの場合を灰色で色付けしている．

病名・医療行為抽出の F 値に着目する．まず，いずれのコーパスの場合も，訓練済み単語分散表現を使うことにより F 値が改善した．Wikipedia という必ずしも医療分野とは関係がない大規模コーパスから学習した訓練済み単語分散表現であっても，病名・医療行為の抽出の性能を向上させることができることが分かる．以下においては単語分散表現を使った結果について考察する．

NTCIR MedNLP コーパスに対して手動で医療行為をアノテートした「MedNLP 手動コーパス」では，病名抽出の F 値が 0.84 であった．少量のコーパスであるが，病名の固有表現のバリエーションがそれほ

表 3　*評価結果. P:精度; R:再現率; F:F 値.*

| コーパス | 単語分散表現 | 病名 | | | 医療行為 | | | 時間 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | P | R | F | P | R | F | P | R | F |
| MedNLP 手動 | 無 | 0.80 | 0.84 | 0.82 | 0.61 | 0.42 | 0.48 | 0.75 | 0.83 | 0.78 |
| | 有 | 0.82 | 0.87 | **0.84** | 0.63 | 0.45 | **0.52** | 0.79 | 0.83 | 0.81 |
| ICD 自動 | 無 | 0.64 | 0.69 | 0.66 | 0.45 | 0.42 | 0.43 | 0.88 | 0.90 | 0.89 |
| | 有 | 0.68 | 0.74 | 0.71 | 0.50 | 0.47 | 0.48 | 0.93 | 0.93 | 0.93 |
| ICD 手動 | 無 | 0.66 | 0.70 | 0.69 | 0.52 | 0.61 | 0.56 | 0.87 | 0.91 | 0.89 |
| | 有 | 0.69 | 0.76 | 0.72 | 0.59 | 0.69 | 0.63 | 0.92 | 0.90 | 0.91 |
| MedNLP 手動＋ ICD 自動 | 無 | 0.73 | 0.78 | 0.76 | 0.49 | 0.46 | 0.47 | 0.84 | 0.87 | 0.85 |
| | 有 | 0.79 | 0.82 | 0.80 | 0.48 | 0.52 | 0.50 | 0.88 | 0.89 | 0.89 |
| MedNLP 手動＋ ICD 手動 | 無 | 0.74 | 0.78 | 0.76 | 0.63 | 0.65 | 0.64 | 0.84 | 0.87 | 0.86 |
| | 有 | 0.78 | 0.82 | **0.79** | 0.62 | 0.72 | **0.67** | 0.89 | 0.89 | 0.89 |

ど大きくないために，比較的高い性能が得られると考えられる．医療行為抽出の F 値は 0.52 と低い．表 1 から分かる通り，医療行為の出現数が少ないため，十分な学習ができなかったと考えられる．

「ICD コーディングトレーニング第 2 版」から自動的に作成した「ICD 自動コーパス」では，病名抽出の F 値が 0.71，医療行為抽出の F 値が 0.48 と，特に医療行為抽出の性能が低い．「ICD 自動コーパス」を手動で修正した「ICD 手動コーパス」では，医療行為抽出の F 値が 0.63 に改善した．本研究で採用した自動アノテーションの方法は確実性が低く，手動による修正が必要であると分かる．

「MedNLP 手動コーパス」に「ICD 手動コーパス」と連結したコーパスでは，「MedNLP 手動コーパス」単独と比べて，病名抽出の性能はやや下がるが，医療行為抽出の F 値が 0.52 から 0.67 に向上した．高い性能とは言えないが，手動で構築した少量のアノテーション付きコーパスと，大規模なアノテーション無しのコーパスから学習した単語分散表現を活用することにより，医療行為抽出の性能を上げることができた．

時間表現抽出の F 値は 0.9 程度の高い性能を示したが，正規表現で自動的にアノテートしたために，限られた種類の表現のみをアノテーションしている可能性がある．データの精査が必要である．

## 結言

本論文では，まず，電子カルテ監査等の医療事務の生産性向上や，医療の質の向上のためには，電子カルテの活用が重要であり，特に，病名だけでなく，手術・処置・検査・投薬といった医療行為を医療文書から抽出することが重要であることを論じた．

以上の観点から，少量の医療文書コーパスながら，病名だけでなく，手術・処置・検査・投薬といった医療行為をアノテートしたコーパスを作成し，コーパスから深層学習による病名と医療行為の固有表現抽出器を学習し，その性能を評価した．評価の結果，アノテーション付き医療コーパスを整備することで，深層学習を使って医療文書から病名だけでなく医療行為を抽出できる可能性が示された．また，Wikipedia という必ずしも医療分野とは関係がない大規模コーパスから学習した訓練済み単語分散表現であっても，病名や医療行為の抽出の性能を向上させることができることが分かった．

今後の課題としては，第一に，本研究で作成した病名・医療行為のアノテーションの精査と，さらにアノテーション付き医療文書を増やすことにより，医療概念抽出の性能を向上させることを目指す．第二に，医学事典等の医療分野の生コーパスを収集し，単語分散表現を教師無し学習で獲得し，医療概念抽出に利用することを予定している．

**謝辞**

本研究は株式会社ユーラスエナジー秋田港より助成を受けたものである. ここに記して謝意を表する.

**文献**

Aramaki, E., Morita, M., Kano, Y., and Ohkuma, T. (2014). Overview of the NTCIR-11 MedNLP-2 Task. In Proc. the 11th NTCIR Workshop Meeting on Evaluation of Information Access Technologies.

荒牧英治, 岡久太郎, 矢野憲, 若宮翔子, 伊藤薫（2017）.「大規模医療コーパス開発に向けて」『言語処理学会 第 23 回年次大会発表論文集』, 1200-1203.

荒牧英治, 若宮翔子, 矢野憲, 永井宥之, 岡久太郎, 伊藤薫（2018）.「病名アノテーションが付与された医療テキスト・コーパスの構築」『自然言語処理』25（1）, 119-152.

Graves, A & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM networks. In Proc. IJCNN.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural Architectures for Named Entity Recognition. In: Proc. NAACL 2016, 260-270.

Morita, M., Kano, Y., Ohkuma, T., Miyabe M. ,& Eiji Aramaki (2013). Overview of the NTCIR-10 MedNLP Task. Proc. 10th NTCIR Conference, 696-701.

佐藤敏紀, 橋本泰一, 奥村学（2017）.「単語分かち書き辞書 mecab-ipadic-NEologd の実装と情報検索における効果的な使用方法の検討」『言語処理学会第 23 回年次大会（NLP2017）』, NLP2017-B6-1.

鈴木正敏, 松田耕史, 関根聡, 岡崎直観, 乾健太郎（2016）.「Wikipedia 記事に対する拡張固有表現ラベルの多重付与」『言語処理学会第 22 回年次大会（NLP2016）』, A5-2.

高間康史, 串間宗夫, 砂山渡（2015）.「TETDM を用いた電子カルテ分析支援ツールの開発と実カルテ分析での検証」『人工知能学会論文誌』30（1）, 372-382.

鳥羽克子（2006）.「ICD コーディングトレーニング第 2 版」, 医学書院.

Uzuner, O. (2008). Second i2b2 workshop on natural language processing challenges for clinical records. Proc. AMIA Annual Symposium, 1252-3.

WHO (1992). ICD10: International Statistical Classification of Diseases and Related Health Problems. World Health Organization. http://www.who.int/classifications/icd/en/ .

山田ひとみ（2017）.「電子カルテ時代における診療録の質に関する研究」, 兵庫県立大学大学院, 博士学位論文, 24506 甲第 319 号.

矢野憲, 伊藤薫, 若宮翔子, 荒牧英治（2017）.「深層学習による医療テキストからの固有表現抽出器の開発とその性能評価」『人工知能学会第 31 全国大会論文集』, 2J2-OS-16a-4.

（2019 年 6 月 30 日受付
2019 年 7 月 9 日受理）

# Extracting Disease Name and Medical Act from Clinical Text Using Deep Learning

Kohji Dohsaka[1], Masaki Ishii[1], Hidekatsu Itoh[1],

[1] *Department of Information and Computer Science, Faculty of Systems Science and Technology, Akita Prefectural University*

In the medical field, electronic medical records are information sources that lead to the improvement of hospital operations and healthcare quality. The analysis of clinical documents can be used to improve the productivity of medical services during electronic medical record audits. In this investigation, we developed a named entity recognizer that extracted disease names and medical interventions from clinical documents. In previous research on medical language processing in Japanese, a corpus had been developed where only disease names were annotated, and based on the corpus, a named entity recognizer for disease names had been developed. However, it is important to recognize not only disease names but also medical interventions such as surgery, treatment, tests, and medication in order to perform an electronic medical record audit. Therefore, we developed a corpus where medical interventions were annotated by constructing a named entity recognizer for disease names and medical interventions. It was shown that it was possible to extract medical interventions from clinical documents by preparing an annotated corpus.

**Keywords:** medical language processing, electronic medical record, medical concept extraction, named entity recognition

Correspondence to Kohji Dohsaka, Department of Information and Computer Science, Faculty of Systems Science and Technology, Akita Prefectural University, 84-4 Ebinokuchi, Tsuchiya, Yurihonjo 015-0055, Japan. E-mail: dohsaka@akita-pu.ac.jp