一般口演

一般口演10
データマイニング・テキストマイニング

2017年11月21日(火) 16:00 ～ 17:30 F会場 (10F 会議室1004-1005)

[2-F-3-OP10-6] オープンエンド型発見手法と深層学習を用いた外来初診時記録からの候補病名予測の検討

山ノ内 祥訓[1], 廣瀬 隼[2], 宇宿 功市郎[2] （1.熊本大学医学部附属病院 総合臨床研究部, 2.熊本大学医学部附属病院 医療情報経営企画部）

【はじめに】医療者が記載する経過記録やレポートは、自然言語で記載されることから二次活用が困難である。当院では2017年1月に稼働した新病院情報管理システムのデータ分析環境の一部に、文章内の意味のある一連のユニット(エンティティ)で分割でき、単語辞書が不要なオープンエンド型発見手法による自然言語解析エンジンを導入した。本研究では、この手法を利用して外来初診時における診療記録を解析し、深層学習することで病名が予測できるかを検証する。
【方法】2015年1月から2017年3月より匿名化した初診日 SOAP記録をオープンエンド型発見手法による自然言語解析を行ったのちにエンティティと対応する病名を抽出した。次に、このデータを深層学習させ、エンティティに対する病名の予測精度を検証した。
【結果】自然言語解析で初診記録52904件が検出され、481446のエンティティ、470の病名が抽出された。これを入力ベルトル1000次元、出力ベルトル470次元に次元圧縮したデータとし、隠れ層4層の多層パーセプトロンを用いて学習させた。実際の病名と予測した病名を比較した結果、全体平均で病名適合率0.95(0.75-1.0)、病名再現率0.61(0.29-1.0)、 AUC0.86(0.65-1.0)、 F値0.74となった。再現率の低い場合には、同義語を含むエンティティが多かった。
【考察】外来初診時の記録をもとに病名を平均95%予測することができた。適合率の高さは、本手法により記載内容の特徴を抽出できた可能性があると考えている。再現率の低さは、過学習であること以外に、同義語同士の組み合わせで特徴量がノイズに紛れたことが要因と考えている。同義語の集約化で精度向上を期待している。今後は予測病名から推定されるエンティティと実際のエンティティの差異を可視化し、記載の支援を行えるアプリケーションの試作を予定している。

オープンエンド型発見手法と深層学習を用いた外来初診時記録からの候補病名予測の検討

山ノ内 祥訓[*1]、廣瀬 隼[*2]、宇宿 功市郎[*2]
*1 熊本大学医学部附属病院 総合臨床研究部、*2 熊本大学医学部附属病院 医療情報経営企画部

Prediction of Diagnosis from first visit records of outpatients Using Open-end discovery method and Deep Learning

Yamanouchi Yoshinori[*1], Hirose Jun[*2], Usuku Koichiro[*3]
*1 Department of Clinical Investigation, Kumamoto University Hospital,
*2 Department of Medical Informatics and Administration Planning, Kumamoto University Hospital,

We have introduced a natural language analyzing system using an open-ended discovery method in conjunction with updating the hospital information system update in January 2017. This new analyzing system has several features, those are dividing sentences into a meaningful series of units (entity) not into a word, and a word dictionary is not essential. In this study, by using this novel technique and a Deep Learning method, we have analyze the medical records whether it is possible to predict diseases using data at the time of first visit in outpatient clinic. The data were extracted from the medical records of first visit outpatient clinic from January 2015 to March 2017. We have collected 51,409 records, and have chosen 462,310 types of entities and 426 types of diseases. To make our analysis easier, we have done dimensional compression of the input vector to the 4,550 from the initial data. Then we have trained compressed data using a Multi-Layer Perceptron of hidden 4 layers. In the result, disease precision rate was 0.936, recall rate 0.704, AUC0.861, and average F value 0.803 in overall, respectively. The low rate of recall is considered to over-fitting data bias and small volume of initial data. We are now aiming to improve the accuracy of the novel system by alternating algorithm and increasing patient data.

Keywords: deep learning, text mining, natural language processing, electronic medical record system

1. 背景

電子カルテシステムでは医療者により日々様々なデータが入力され蓄積されている。これらのデータには、オーダ情報や検査結果など構造化されたデータもあれば、経過記録や検査所見などの自然言語や画像、音声といった構造化されていないデータも存在する。これら蓄積されたデータを二次活用しようという取り組みは以前よりなされているが、その多くは構造化されたデータが対象である。

構造化されたデータは、格納される内容がマスタなどを利用してコード化されているため、多くの検索・分析ツールを利用して容易に可視化し分析することが可能である。一方、非構造化データは格納されているデータの形式が決まっておらず、フリーテキストデータや画像データ、音声データなどどのような形式で格納されるかはその時に応じて異なることが多い。そのため、単純な検索・分析ツールでは取り扱うことが非常に困難である[1)]。

非構造化データのうち、自然言語データは電子カルテシステムでは多く記録されているデータであるが、このデータを分析するためにはまず自然言語処理を用いて自然言語自体を解析する必要がある。日本語における一般的な自然言語処理は、最初に形態素解析で単語分割を行い、そこから単語間の関連を解析する「係り受け解析」などの構文解析を経て、類語辞典や固有辞書を用いた意味解析を行う[2)]。自然言語処理のアルゴリズム自体も複雑であるが、加えて解析を行う場合必要なのが単語辞書や類語辞典である。しかし、日々新しい単語や表現、概念が生まれてくる現状において、常にこれらの更新を行うことは困難である[3)]。

当病院では、2017 年 1 月に新病院情報管理システムの更改に合わせてデータ分析環境も更新した。このデータ分析環境には、これまで存在した診療データウェアハウス(DWH)機能や Business Intelligence（BI）機能に加え、前述の自然言語処理の弱点を改善したオープンエンド型発見手法を用いた自然言語解析機能が新しく導入された。

これにより、従来では困難であった自然言語解析を行うことが期待できることから、本手法を用いた自然言語データ分析の有用性の確認と深層学習を利用した予測手法の検討を行った。

1.1 オープンエンド型発見手法「iKnow」について

オープンエンド型発見手法は、これまでの自然言語解析と異なり分割の単位を単語ではなくエンティティと呼ばれる文章内の意味のある一連のユニットで文章を分割することができるものである[4)]。エンティティとは、意味的なつながりの強い連続した単語からなっており 1 つの単語よりも文章の特徴を表現しやすい(図 1)。また、文章中の出現状況が近いエンティティを抽出できるほか、構文解析により元の文章の語順が多少異なっていても整理された語順にエンティティの順番に並べ替える機能を有している。さらに、解析のための辞書を有しておらず解析元となる文章から必要となる情報を抽出するボトムアップ形式となっているため、事前の辞書作成や運用時のメンテナンスが不要である。

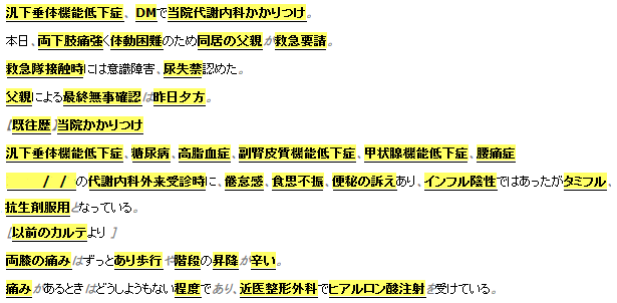

図 1 文章を iKnow で分割したときの例

本研究で使用したオープンエンド型自然言語解析エンジンは InterSystems 社の iKnow[5]である。iKnow では解析処理を行うとエンティティごとの指標として頻度、スプレッド、ドミナンスという 3 つの指標が算出される。また、エンティティ同士の指標としてパスシーケンスと近接度が算出される。各指標は以下の通りの意味を持っている。

- 頻度
  文書全体でそのエンティティが出現した回数を示す。
- スプレッド
  文書全体でそのエンティティが出現した文書数を示す。
- ドミナンス
  文書全体もしくはある 1 つの文書でそのエンティティの構文上の位置や頻度、等の値をもとに算出され、文章中における重要度を示す。
- 近接度
  文書全体でエンティティ間のパスシーケンスをもとに算出され、文書全体におけるエンティティ同士の関連性を示す。
- パスシーケンス
  ある 1 つの文書の内容を構文上整理された状態に変換したときのエンティティの出現順序を示す。

これらの指標を組み合わせることによって文章内における特徴を定量的に示すことが可能となり自然言語処理の精度向上が期待できる。

1.2 深層学習について

深層学習(Deep Learning)は、機械学習アルゴリズムの一つである[5]。機械学習には様々なアルゴリズムが存在するが、その中でニューラルネットワークという生物の脳神経ネットワークを参考にモデル化したアルゴリズムを多層化したものである(図 2)。ニューラルネットワークは入力層、中間層、出力層の 3 つに分かれており各層のノードに対して前後の層からデータ伝達のためのパスが接続される。このパスにはそれぞれ係数(重み)を設定することができ、この係数によってデータの強弱が決定される。また、ノードは活性化関数と呼ばれる関数で構成されており、ここで入力された重みづけデータを活性化関数で処理しその結果を出力する。ニューラルネットワークでいう学習とは、この各ノード間のパスの係数を調整することである。

深層学習は、このニューラルネットワークでは 1 つしかなかった中間層を複数重ねることにより、ネットワーク内の自由度を高めることで複雑なモデルを表現することが可能となった。近年、アルゴリズムの改善やコンピュータの性能向上により画像認識やオートエンコーダと呼ばれる特徴抽出器など様々な分野で高い精度を発揮している。

本研究で使用した深層学習環境は、Google 社で開発され現在オープンソースで開発が進められている TensorFlow[7]と、TensorFlow のフロントエンドフレームワークとしてこちらもオープンソースで開発が進められている Keras[8]である。

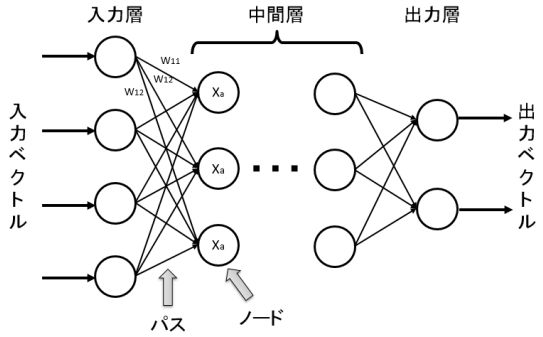

図 2 ニューラルネットワークの構成

2. 目的

本研究では、外来初診時の診療記録をオープンエンド型発見手法により自然言語解析しエンティティを抽出して、その結果を深層学習の入力とすることで、初診時の病名が予測できるかを検証することにある。

3. 方法

3.1 対象となるデータ

2015 年 1 月 1 日から 2017 年 3 月 31 日までに当病院の電子カルテシステムに登録された外来初診患者を対象とした。抽出された患者数は延べ 79,321 名であった。対象となる診療記録は全ての医療者が記載した診療記録のうち初診時記録及び経過記録の患者の訴え(Subjective findings, S)と客観所見(Objective findings, O)の記録を使用した。これを、オープンエンド型発見手法を使用した言語解析エンジンに入力し、3,635,235 件のエンティティを抽出した。データラベルとなる病名については外来初診日に有効な主病名の ICD10 コードから DPC で使用されている 6 桁の疾患コードに変換したものを使用した。抽出された病名は 470 件であった。患者の氏名や性別、生年月日、年齢、診療日といった個人を特定できる情報や背景情報は抽出せず、患者 ID などキー情報となるデータについては乱数による鍵付きハッシュ関数を用いた不可逆処理を行い、匿名化処理を施した。

3.2 データセットへの変換

抽出されたデータを元に深層学習で使用するデータセットへ変換した(図 3)。入力ベクトルとなる診療記録は、診療記録 1 件についてあるエンティティが出現したら"1"、しなかったら"0"となるように 2 値化したスパース行列として変換した。しかし、抽出されたエンティティは 3,635,235 件という膨大なデータとなるため、必要なエンティティを選択する必要が生じた。まず、数値データのみが含まれているエンティティなど一部のエンティティを除外した。次に、ドミナンスの大きい順に上位 20%のみを抽出し、後のエンティティは全て除外した。最後に、5 件未満のエンティティしかなかった診療記録はレアケースとして除外した。よって解析対象のエンティティは 462,310 件で、入力ベクトルは 462,310 次元となった(表 1)。

出力ベクトルとなる病名は、入力ベクトルと同様にある疾患が出現したら"1"、しなかったら"0"となるように 2 値化したスパ

ース行列として変換した。こちらは全データのうち 3 件未満しか出現しなかった病名はレアケースとして除外した。よって出力病名は 426 件、ベクトルは 426 次元となった。このデータ変換によりデータセットの件数は 51,409 件となった。

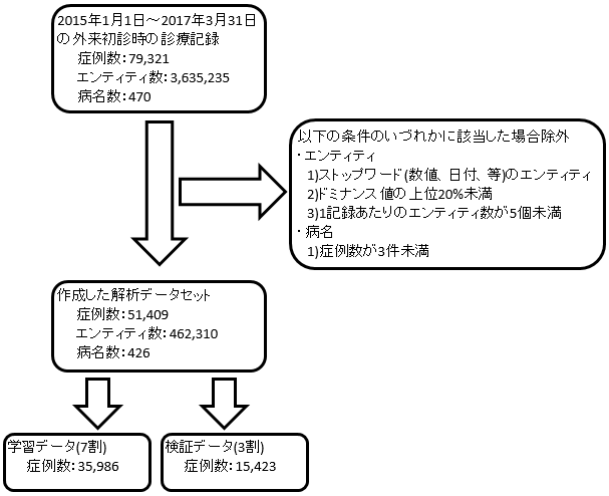

図 3 使用するデータの選択

表 1 使用するデータセットのエンティティ内訳

入力データ	
診療記録区分名	エンティティ件数
初診入院時記録・主訴	26,044
初診入院時記録・現病歴	80,478
初診入院時記録・身体所見	26,510
初診入院時記録・検査所見	25,563
経過記録・Subjective	200,115
経過記録・Objective	103,600
合計	462,310

使用した環境は、院内の病院情報ネットワークに設置されたサーバ上(CPU 8 個、メモリ 64GB)で実行し、プログラム開発は Visual Studio 2015 上で C#を使用して行った。

3.3 深層学習での学習

本研究で使用した深層学習の構成は図 4 の通りである。入力ベクトルは、診療記録区分ごとに分割し、それぞれ 1 層のオートエンコーダを使用した特徴抽出を行い 100 分の 1 に次元圧縮した後、圧縮したベクトルを全て結合した 4,550 次元とした。出力ベクトルは、病名データである 426 次元をそのまま使用した。深層学習は、中間層 4 層の多層パーセプトロン(Multi-Layer Perceptron:MLP)を用いてマルチラベル分類問題として構成した。使用したオプティマイザは adam、loss 関数は binary clossentropy、メトリックは accuracy を指定した。学習は、バッチサイズ 512 データで 500 エポックから開始し、中間層のノード数を 500～4,000 の範囲を 500 刻み、各層のドロップアウト率を 0～0.5 の範囲を 0.1 刻みで変化させチューニング(grid search)を行った。それ以外のハイパーパラメータは初期値のままとした。学習データは全データの 7 割となる 35,986 件、検証データは全データの 3 割となる 15,423 件とし、学習データと検証データの選択はランダムに行った。

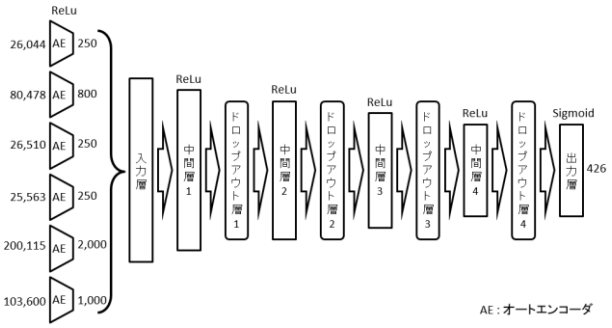

図 4 深層学習モデルの構成

実行環境は、Amazon 社が提供するクラウドサービスである Amazon Web Service(AWS)の仮想マシンである Elastic Compute Cloud(EC2)を使用した。EC2 インスタンスは、c4.8xlarge(vCPU 36 個、メモリ 60GB)、m4.10xlarge(vCPU 40 個、メモリ 160GB)、p2.2xlarge(vCPU 4 個、NVIDIA K80 GPU 1 個、メモリ 61GB、GPU メモリ 12GB)の 3 種類をデータ量と目的に応じて使い分けた。なお、すべてのインスタンスは AWS 内に構築した仮想プライベートクラウド(VPC)内に設置し、ストレージは全てブロックレベルで暗号化したうえで、院内とのデータのやり取りは全て SSL/TLS による暗号化通信にて行った。また、深層学習実行サーバは直接インターネットとの通信ができないように設定した。プログラム開発は Python バージョン 3.5.2、TensorFlow のバージョンは 0.12.1、Keras のバージョンは 1.2.1 の環境で行った。

3.4 評価方法

作成した深層学習のモデルの評価は、全データのうち 3 割となる 15,423 件のデータについて病名の予測処理を行い、各例において予測された病名の予測確率が 0.5 以上であれば病名ありとして数え、0.5 未満であれば病名なしとして数えて、予測データを生成した。この予測データと実際の病名データを比較し、適合率(precision)、再現率(recall)、AUC、F 値(調和平均)の 4 つの指標を算出して評価した。

4. 結果

深層学習モデルのハイパーパラメータチューニングの結果、決定したパラメータは表 2,3 の通りとなった。決定したモデルの学習データの loss データと検証データの loss データの推移を図 5 に示す。モデル自体は 5,000 エポックまで実行したが、学習データの loss 値はエポックが進むにつれて低下し 2,000 エポックを超えた付近でほぼ一定の値となった。しかし、検証データの loss 値は一度 60 エポック付近で一旦低下した後に上昇し、その後学習データと同じく 2,000 エポックを超えた付近でほぼ一定の値となった。また、学習データと検証データの loss 値の差が 11.4 倍だった。

表 2 使用したハイパーパラメータ(中間層)

中間層	
層名称	ノード数
中間層1	4,000
中間層2	3,000
中間層3	500
中間層4	500

表 3 使用したハイパーパラメータ(ドロップアウト層)

ドロップアウト層	
層名称	ドロップアウト率
ドロップアウト層1	0.5
ドロップアウト層2	0.2
ドロップアウト層3	0.2
ドロップアウト層4	0.2

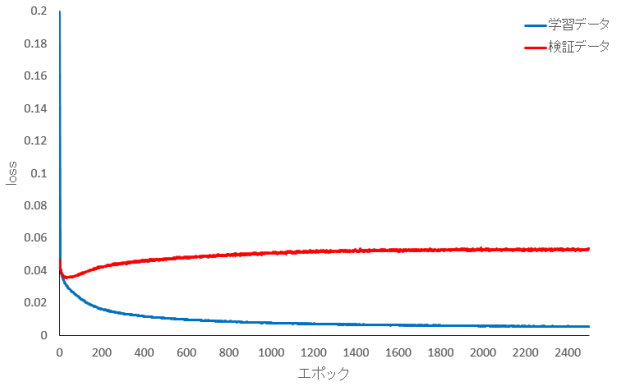

図 5 MLP の学習曲線

検証データを使用した病名予測結果は、全 426 病名の平均適合率が 0.936、平均再現率 0.704、平均 AUC0.861、平均 F 値 0.803 となった(表 4)。病名毎に分布を確認すると、適合率はほとんどが 0.9 以上となっているが一部の病名で 0 に近い値となっていて予測できていないことが分かった(図 6)。また、再現率では 0.6～0.7 付近に集中しており、0.9 以上の値となったものもみられるが、一部の病名で 0 に近い値となっていた(図 7)。

表 4 病名予測結果の概要

	平均値	中央値
適合率	0.936	1.000
再現率	0.704	0.724
AUC	0.861	0.864
F値	0.803	0.840

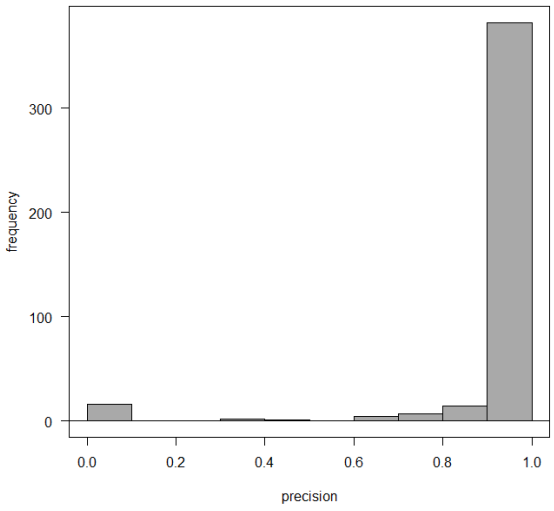

図 6 病名毎の適合率の分布

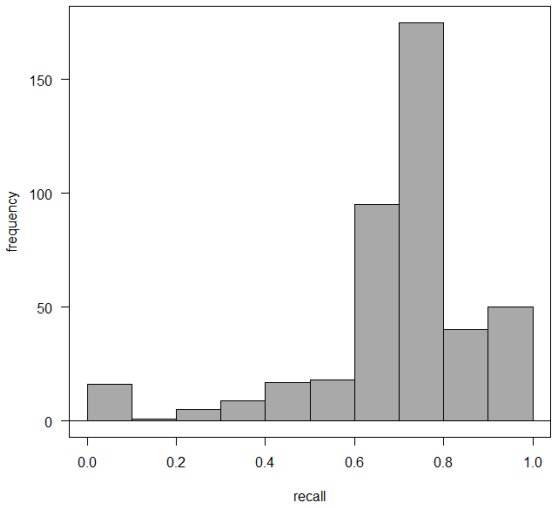

図 7 病名毎の再現率の分布

本研究で使用したデータにおいて病名の出現頻度と予測結果を確認した。検証データ中に出現した病名は 418 個であり、8 個の病名が検証できていなかった。また、病名毎の予測結果を確認すると出現回数が 50 回以下になると F 値がばらつき始めていた(図 8)。データセット全体(51,409 件)における病名の出現回数は最も多い病名が心不全で 9,366 件であり、最も少ない病名は網膜の悪性新生物で 3 件だった。

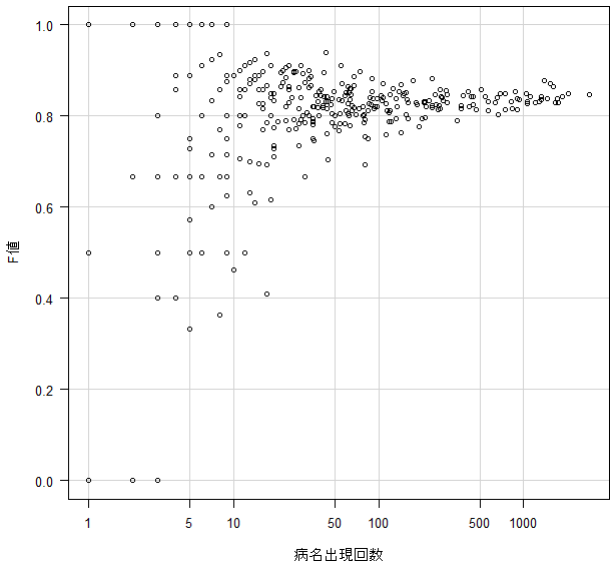

図 8 検証データにおける病名出現回数と F 値の関係

実際の病名と予測された病名の上位 10 位の平均予測確率について、急性心筋梗塞とインフルエンザを例として示す(表 5,6)。尚、この表においては予測確率 0.5 以下の病名も提示するが、適合率、再現率の計算には用いていない。急性心筋梗塞について実際の病名は 2039 症例に対し、予測された病名の平均予測確率は「急性心筋梗塞」が 0.768、「心不全」が 0.571、「静脈炎および血栓(性)静脈炎」が 0.36 となった。出現回数では、2039 症例中「急性心筋梗塞」が 2038 症例、「心不全」が 1916 症例、「静脈炎および血栓(性)静脈炎」が

1661 症例だった。インフルエンザについて実際の病名は 244 症例に対し、予測された病名の平均予測確率は「インフルエンザウィルスが分離されたインフルエンザ」が 0.743、「百日咳菌による百日咳」が 0.148、「その他の凝固障害」が 0.129 となった。出現回数では、244 症例中「インフルエンザウィルスが分離されたインフルエンザ」が 244 症例、「百日咳菌による百日咳」が 170 症例、「その他の凝固障害」が 152 症例だった。

表 5 急性心筋梗塞症例に対する予測された病名

予測病名	平均予測確率	出現回数
急性心筋梗塞	0.768	2,038
心不全	0.571	1,916
静脈炎および血栓(性)静脈炎	0.360	1,661
大動脈のアテローム＜じゅく＜粥＞状＞硬化(症)	0.296	1,585
先天性ヨード欠乏症候群	0.240	1,407
狭心症	0.223	1,591
無症候性ヨード欠乏性甲状腺機能低下症	0.194	1,570
その他の凝固障害	0.177	1,433
急性A型肝炎，肝性昏睡を伴うもの	0.176	1,128
先天梅毒	0.174	1,055

表 6 インフルエンザ症例に対する予測された病名

予測病名	平均予測確率	出現回数
インフルエンザウイルスが分離されたインフルエンザ	0.743	244
百日咳菌による百日咳	0.148	170
その他の凝固障害	0.129	152
サルコイドーシス	0.120	143
急性心筋梗塞	0.103	142
心不全	0.103	157
急性アメーバ赤痢	0.093	150
サルモネラ敗血症	0.084	146
歯の発育および萌出の障害	0.084	125
その他の反応性関節障害　肩甲帯	0.074	106

各病名における出現エンティティは、全体のエンティティ出現順位より、実際にその病名である場合においてエンティティの出現順位が 50 位以上繰り上がったエンティティのうち上位 20 位を S,O それぞれ抽出した。出現頻度の繰り上がったエンティティは当該病名の場合に記載されると考えられるエンティティが多く含まれていた。例として、急性心筋梗塞とインフルエンザに対して特徴的なエンティティを示す(表 7,8) 。

表 7 急性心筋梗塞で特徴的なエンティティ

経過記録･S			経過記録･O		
エンティティ	病名順位	全体順位	エンティティ	病名順位	全体順位
__時間	20	71	ecg	9	59
手術時間	26	119	__mmhg	11	62
胸痛	30	124	正常範囲内	12	69
息切れ	43	126	左室径	14	85
出血量	44	139	呼吸性変動良好	17	92
全身麻酔	46	177	開放良好	19	95
動悸	47	123	心エコー	22	76
常用薬	55	214	右房圧	25	124
耐術能	57	239	洞調律	28	98
全身麻酔下	60	143	右心系の拡大なし	30	161
術前検査	67	208	壁厚	31	132
心電図	70	197	lung	33	89
心エコー	71	156	左室壁運動良好	36	173
呼吸器外科	73	222	ctr __%	39	180
かかりつけ	81	146	ivc	40	145
術式	83	230	拡大	42	157
心房細動	88	190	abi	44	222
adl自立	89	397	推定肺動脈圧	45	239
胸部症状	92	261	右心系の拡大	46	223
階段	97	248	弁逆流	49	181

表 8 インフルエンザで特徴的なエンティティ

経過記録･S			経過記録･O		
エンティティ	病名順位	全体順位	エンティティ	病名順位	全体順位
発熱	1	63	陰性	16	134
今朝	18	421	発熱	20	118
経過	19	87	改善傾向	25	169
術後	20	97	__ml	26	83
倦怠感	23	157	著変なし	27	109
きつい	25	94	bt __℃	31	150
頭痛	29	88	room air	38	266
良く	31	102	呼吸	44	160
__℃	33	830	左右差なし	52	131
トイレ	34	115	__度	57	318
嘔気	41	218	主治医	60	152
下痢	47	163	感覚	61	218
悪寒	54	1812	発赤	62	192
悪く	55	136	入院	64	144
ステロイド	57	201	経過観察	65	133
自宅	58	118	neut	66	247
間質性肺炎	62	227	rr	67	234
はい	63	140	状態	73	203
お腹	64	144	浸潤影	77	524
腎機能	65	277	経過	78	143

5. 考察

オープンエンド型発見手法を利用して初診時記録からエンティティを抽出し、それを深層学習で処理することで病名の予測ができることを確認した。予測の精度は、平均適合率が 93.6%、平均再現率が 70.4%となった。再現率は先行研究と比較して高い値ではないが、今回の検討では、初診時において自然言語で記録される背景情報や主訴、所見の記録を利用して病名の候補を予測できる可能性が示唆された。

モデルの学習結果では、今回チューニングで変化させたパラメータ以外にも層の数やバッチサイズ、学習率をはじめとする様々なパラメータを調整する必要があるが、探索に非常に時間がかかる(今回は 500 エポックの学習で 1 モデル約 2 時間)ことからパラメータを限定してチューニングを行った。学習の推移を確認すると、学習時の loss データは学習データと検証データで loss 値が約 10 倍異なることから未知データに対応できない過学習状態ではないかと考えている。

また、使用したモデルは、オートエンコーダと MLP の二つを組み合わせた構成とした。前段のオートエンコーダは、入力ベクトルが約 360 万次元に対して症例数が約 5 万件というデータの不均衡を解消するため、特徴量抽出の次元圧縮を役割とした。他の次元圧縮の手段として主成分分析(PCA)や潜在的ディリクレ配分法(LDA)などが存在するが、実際にデータを投入して実行したところリソース不足で処理が終わらず圧縮が行えなかったことから、オートエンコーダを使用した。後段のモデルは MLP を使用したが、深層学習には他に有名なアルゴリズムとして畳み込みニューラルネットワーク(CNN)がある。CNN は画像認識などで高い精度を出すことが知られているが、ベクトルの各要素の値だけでなく周辺の値も使って演算を行うため、ベクトル上に配置するエンティティの順番で精度が大きく変化することから、本研究では使用しなかった。

検証データを使用した病名の予測結果では、適合率は、426 病名中 383 病名が 0.9 以上の適合率だった。ただ、適合率が 1.0 となったのが 241 病名あり、そのうち検証データ中の病名出現回数が 10 回以下だった病名が 118 件含まれていた。また、適合率が 0.1 以下の病名は 16 病名であるが全て 0.0 であり、そのうち出現回数が 0 回の病名が 8 件で、残りの 8 件も 3 回以下の出現回数だった。このことは、データラベルとなる病名の出現回数が少ないと、レアケースとして無視されるケースと、少数のパターンで学習してしまうケースの二通りがあったと考える。再現率は、平均として 70.4%であり、病名毎の

再現率分布においても 0.6～0.7 に値が集中していることから、本来の病名のうち 6～7 割程度しか予測できていない。また、こちらも再現率が 1.0 の病名が 49 件あり出現回数は全て 9 回以下で、再現率が 0.0 の病名は 16 件あり、そのうち出現回数が 0 回の病名が 8 件で、残りの 8 件も 3 回以下の出現回数だった。また、エンティティを確認したところ、「嘔気」や「アレルギー・喘息」といった単語だけで切り出され、有無などの状態が含まれていないエンティティが多く見られた。他にも、「ふらつきなし」「ふらつき無い」「ふらつき-」といった同一の意味を持つが表現の異なるエンティティが散見された。こういった単語や類似表現が入力データ上別の次元として出現することで、特徴として抽出できるだけの情報量を獲得することができなかったことも再現率低下の原因であると考える。

病名の予測結果と病名の出現回数について予測結果の F 値の分布を確認すると、病名の出現回数が 50 回以下になると F 値が平均値から外れ始めたことが分かり、10 回以下になると F 値が 0.0 や 1.0 といった予測できているとは考えにくい値がみられるようになった。最も多い病名は心不全で、データセット全体で 9,366 件、検証データで 2,858 件であった。最も少ない病名は網膜の悪性新生物で、データセット全体で 3 件、検証データで 0 件だった。機械学習は頻度の多いデータに予測が引きずられる傾向があり、本研究においても同様の現象が見られた。データラベルとなる病名の出現回数は最大値と最小値の差が 1 桁程度になるように調整が必要と考える。また、実際の病名と予測した病名の関係について確認したが、こちらは適合率の通り、ほぼ実際の病名と同じ病名が高い平均予測確率となっており、それ以外の病名についても実際の病名に近い病名を予測していることが多いことから、もともと症例数の多い病名の予測は行えていると考える。

病名毎に特徴的なエンティティの確認は、エンティティの出現回数で見ると大きく異なることから出現回数の順位変動を指標として考えた。本研究では全体の出現順位より病名毎の出現順位が 50 位以上繰り上がったエンティティの中から、病名毎のエンティティの出現順位上位 20 位までを抽出して確認した。抽出したエンティティの多くは、その病名が付いたときに記載されると考えられるエンティティであった。このことから、病名毎に特徴のある表現を抽出する 1 手段として有用である。しかし、エンティティの内容としては単語が多かった。これについていくつかの記載内容を確認したところ、診療記録には文章として記述される場合と、箇条書きなど単語の羅列として記述される場合、テンプレート等を使用したため単語と区切り文字で区切られた記述となる場合があった。また、記載者によって、文章として記載している人もいれば単語を並べただけの人もおり、記載レベルがまちまちであった。これらのデータを区別なくオープンエンド型発見手法による解析を行ったことで、アルゴリズムが単語に近い部分で切り出したものと考えられる。これについては、iKnow のチューニングや記載内容のパターンを考慮した解析を行う必要があると考える。

本研究の結果、データの質向上と過学習対策が課題として挙げられた。まず、データの質向上では作成したデータセットの見直しである。本研究では、オープンエンド型発見手法で抽出されたエンティティのうちドミナンス値の上位 20%のみを深層学習に使用した。しかし、使用したエンティティの内容を確認すると単語とほぼ同じ粒度のエンティティが多く抽出されていた。一方、使用しなかったドミナンス値の低いエンティティでは、複数の単語の組み合わせとなっているエンティティが多かった。また、類義語も多かったことから、エンティティの切り出しとエンティティの構文上の指標の提供のみでは、言語解析の精度向上に不十分であることが分かった。エンティティ自体を類語辞典に自動分析することなどを今後検討すべきと思われる。また、各病名の特徴をとらえたエンティティを選択する手段を開発する必要があるともいえる。病名については、ICD10 の粒度では分類する数が多くなり精度が低下する可能性があったことから DPC 疾患コードに変換して解析を行った。実用性の面では問題があることも否定できないが、今回検討した症例数やデータ数からは妥当と考えている。

次に、過学習対策としては、データの質が要因であるとともにデータ量の不足も大きいと考えている。対象とした検索期間は 2 年 3 か月と短くないが最終的な症例数は約 5 万件であり、これは深層学習の学習データとして不足しており、さらに当院の病院特有のモデルになっていることも考えられた。そのため、他病院のデータを含めたモデルを作成することでより一般的な性能を発揮するモデルが作成できると考える。また、病名の偏りが大きく予測結果に影響を与えていた問題は、件数の多い病名のデータをランダムに間引く、少数のデータを複製や組み合わせによりデータ数を増やす、病名のデータ数に応じてモデルを分ける、などの対応が考えられる。以上の対応を行うことにより、過学習にならず安定した精度で病名の予測が可能になれば、これを電子カルテシステムと連動させることで、病名の記載補助や診療記録の質的監査の補助に利用できることを目指せるものと考えている。

参考文献

1) Dongxiao Gua, Jingjing Li, Xingguo Li, Changyong Lianga. Visualizing the knowledge structure and evolution of big data research in healthcare informatics. International Journal of Medical Informatics 2017 ; 98 22-32.
2) 山下貴範, Flanagan Brendan, 若田好史,他. テキストマイニングで抽出された重要単語を利用した重要文の評価. 第 35 回医療情報学連合大会論文集 2015.
3) Demner-Fushman, Dina, Wendy W. Chapman, Clement J. McDonald. What can natural language processing do for clinical decision support?. Journal of biomedical informatics 2009 ; 42 : 5 760-772.
4) 本多正幸,松本武浩,浅田眞瑞,他. 自然言語解析ツール「iKnow」を用いた退院時サマリーデータの評価－長崎大学病院DWHを対象とした取組－. 第 36 回医療情報学連合大会論文集 2016.
5) InterSystems Corporation. iKnow | Embedded Technologies | InterSystems. [http://www.intersystems.com/our-products/embedded-technologies/iknow/ (cited 2017-Sep-01)]
6) Jurgen Schmidhuber. Deep learning in neural networks: An overview. Neural Networks 2015 ; 61 85-117.
7) TensorFlow. [https://www.tensorflow.org/ (cited 2017-Sep-01)].
8) Keras. [https://keras.io/ (cited 2017-Sep-01)].