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

Development and Performance Evaluation of Deep Learning Method for Extraction of Named Entities from Medical Case Reports

矢野　憲[*1]　伊藤　薫[*1]　若宮　翔子[*1]　荒牧　英治[*1]
Ken Yano　Kaoru Ito　Shoko Wakamiya　Eiji Aramaki

[*1] 奈良先端科学技術大学院大学
NARA Institute of Science and Technology

This paper proposes a character based Bi-LSTM-CRM for named entity extraction from Japanese medical case reports and evaluates its performance against previous CRF based approaches.

1. はじめに

電子カルテの普及ならびに人工知能の発展に伴い，患者や疾患の情報が記録された膨大な医療テキストを診断支援などに利活用することへの期待が高まっている．しかし，医療文書は，非文法的かつ断片化した表現を用いて記述されており，処理が難しい．このため，医療テキスト処理に特化した言語処理が注目されている [Chapman 11]．英語の医療テキストの解析には，MAYO Clinic の語彙サーバー，MedLEE (Medical Language Extraction and Encoding system) [Friedman 95]，cTAKES (clinical Text Analysis Knowledge Extraction System) [Ctakes] などのツールがすでに公開され，広く利用されている．一方で，日本語の医療テキストを扱う標準的な言語処理ツールは未だ存在しない．そこで本研究では，日本語の医療テキスト解析ツールの開発を目指す．

医療テキストの解析の主なタスクは 2 つある．1 つ目は医療テキスト中の病名，疾患名を識別（以降，事象認識 **(ER)** と呼ぶ）するタスクである．もう 1 つは，患者が実際に患っている疾患（陽性所見，**P**: **Positive**）とそれ以外の所見（陰性所見，**N**: **Negative**）を分類（以降，**P/N** 分類と呼ぶ）するタスクである．我々はこれまでの研究 [矢野 17] において，医療テキストの特徴を考慮して，これら 2 つのタスクを同時に行い，文字ベースで処理する事象抽出器（以降，文字ベース系列ラベリングと呼ぶ）を構築している（図 1）．この文字ベース系列ラベリングは，形態素解析などの追加システムなしで事象認識することができ，1 回のラベリングで 2 つの処理を行うため，処理速度が速く，実装も容易という利点がある．一方で，ウィンドウサイズや素性の組み合わせなどを決定する素性テンプレートをデザインする必要があるなど，経験則に頼らざるを得ないという問題点がある．

本稿では，深層学習を用いた文字ベースの “ end-to-end” 事象認識器を提案する．これにより，特に文字ベース系列ラベリングで問題となっていたウィンドウサイズの指定が不要となり，文字ベース系列ラベリングの利点のみを継承した事象認識が可能となる．提案する事象認識器の特徴を以下に示す．

文字ベース系列ラベリング：従来の言語処理研究では，単語（形態素）を最小単位とみなすものが多い（図 1 (c)）．しかし，医療テキストは，長く複雑な複合名詞（例えば，「傍大動脈リンパ節郭清」など）や，一般的には用いられないひらがなを多く含む専門用語（例えば，「びまん性」など）が多く出現し，しばしば形態素解析の誤りにつながる．このような医療テキストの特徴を考慮し，文字ベースの処理を行う（図 1 (b)）．さらに，ER タスクと P/N 分類タスクにおいて必要な情報が共通していることが多い．例えば，「～が認められる」「～が認められない」は，ともに病名認識の大きな手がかりであるとともに，P/N 分類の手がかりにもなる．そのため，通常段階的に行う 2 つのタスクを 1 つに融合する．

連絡先：矢野　憲，NAIST, yanoken@is.naist.jp

(a)　入力テキスト

腫瘤は肝細胞癌ではなく形質細胞腫と診断された．

(b)　文字ベース系列ラベリング

腫	瘤	は	肝	細	胞	癌	で	は	な	く	形	質	細	胞	腫	と	診	断	さ	れ	た
O	O	O	B-N	I-N	I-N	I-N	O	O	O	O	B-P	I-P	I-P	I-P	I-P	O	O	O	O	O	O

(c)　単語ベース系列ラベリング

腫瘤	は	肝	細胞	癌	で	は	なく	形質	細胞腫	と	診断	さ	れ	た
O	O	B-N	I-N	I-N	O	O	O	B-P	I-P	O	O	O	O	O

(d)　出力テキスト

腫瘤は<N>肝細胞癌</N>ではなく<P>形質細胞腫</P>と診断された．

図1. 2種類の系列ラベリングにおけるシーケンス表現．B, I, O ラベルは系列ラベリングで推定された入力テキストシーケンス上での固有表現の開始，内側，それ以外の外側をそれぞれ表す．(b) は文字ベース系列ラベリング（提案手法）の結果，(c) は単語ベース系列ラベリング（既存手法）の結果である．

深層学習による“end-to-end”手法の適用：文字ベース系列ラベリングに深層学習を用いる．深層学習の採用には様々な利点がある．まず，医療テキストから固有表現を抽出するために有効な特徴量を，経験則に頼らず自然に学習することができる点である．次に，再帰的ニューラルネットワークを用いることで，前後のコンテキストに依存した系列ラベリングが可能となり，P/N 分

類におけるモダリティの判定に優位に働く点である. CRF による既存手法[矢野 17] でも前後のコンテキストに依存した系列ラベリングが可能だが, コンテキストが予め固定されたウィンドウサイズに限定されるという制約があった. 最後は, より大きな学習データを用いることで, 飛躍的な性能向上を期待できる点である.

実験では, アノテーションが付与された約 500 件の医療症例報告をコーパスとして用いて, 10 分割の交差検証により性能評価を行った. その結果, これまでの文字ベース CRF と比べて $F_{\beta=1}$尺度で, <P>タグ検出については約+5, <N>タグ検出については約+10 の精度向上を確認した. また, 表層の文字以外に, ICD コードや文字種別を追加の文字特徴量として用いることで, 学習回数の初期段階で性能が飛躍的に向上することを確認したが, 学習が進んだ段階ではその効果はほとんど無視できることも確認できた.

2. 関連研究

再帰的ニューラルネットワーク(RNN)による深層学習を系列ラベリングの問題に適応する多くの研究例が報告されている[Chi 16][Lample 16]. RNN はネットワークの一部が再帰的に自身に帰還する構造を持っているため, データの記憶が可能となっている. このため, テキストの系列ラベリングにおいて, 前後のコンテキストの記憶を利用した学習が可能になる. ただし一般的に RNN では, 誤差逆伝搬時の勾配爆発・消失のため, 比較的長いコンテキストの記憶を行うのが困難であった. そのため, LSTM(Long Short Term Memory)が考案された. LSTM はシグモイド関数で定義される忘却, 入力, 出力のゲートの開閉を巧みに制御することで短期・長期の記憶を可能する構造をもっている[Hochreiter 97].

3. 手法

3.1 文字ベース Bi-LSTM+CRF

本研究では, 双方向(前方, 後方)LSTM と CRF を出力層に用いた深層ニューラルネットワークによる系列ラベリングを提案する. 図 2 にそのネットワークの構成を示す. 各時間ステップにおける入力は文字であり, 文字ベクトル変換層により文字辞書から得られた文字 ID を文字ベクトルに変換する. 表層の入力文字以外に, オプションとして事前処理にて文字毎に付与された ICD コード情報と文字タイプ情報を文字ベクトルに追加することも可能である. これらはそれぞれの辞書から得られる特徴 ID を文字と同様にベクトルに変換した文字毎の特徴量である. この連結されたベクトルが双方向 LSTM ネットワークへの入力となる. 次に前方, 後方 LSTM からの出力を結合したものが全結合の隠れ層に入力される. 隠れ層の出力は, BIO ラベルに対応した多クラスの確率情報を与える. シーケンス全体の確率情報が CRF に入力され, 確率的に尤もらしい BIO 系列ラベリングが出力される.

表 1 に入力として表層文字のみを用いた場合の各ネットワーク層を構成する要素のパラメータの次元数を示す. これらの値は, 事前調査により最適な値に調整された.

3.2 ER と P/N 分類

本研究での入力は文であり, 出力は<P>タグや<N>タグが付与された文である. この出力は, (1) ER と (2) P/N 分類の 2 つの処理の結果であり, 処理(1) の結果はタグ付けされた範囲として, 処理(2) の結果は, タグのタイプ(<P>または<N>)として表される. これまでの系列ラベリングでは通常, P/N 分類などのモダリティ

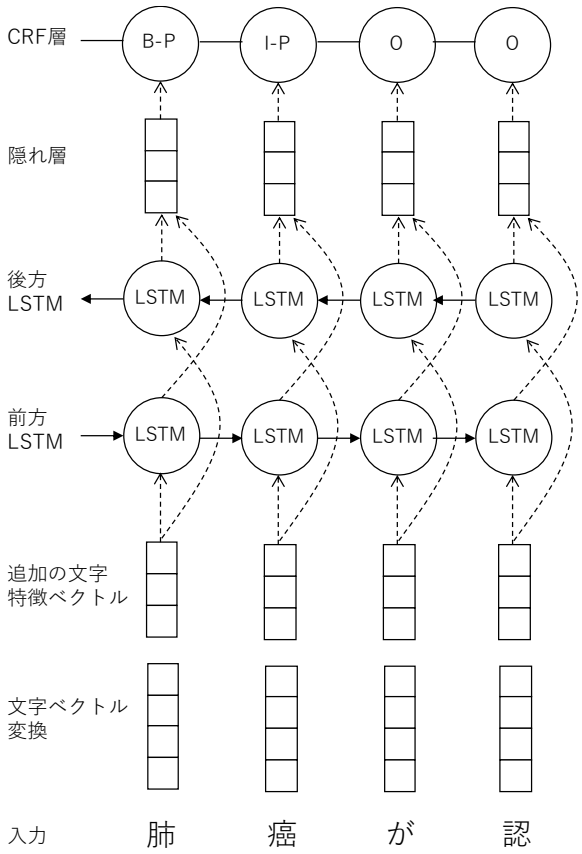


図 2. Bi-LSTM+CRF 系列ラベリングネットワークの構成. 文字単位の特徴ベクトルに変換された時系列文字データは, それぞれ別々に前方, 後方 LSTM に入力される. この出力が結合されたデータは隠れ層を通じて, CRF 層に入力される.

判断は ER の結果に適用して行っていた. これに対し, 提案手法では, "end-to-end"により, 入力テキストからの ER と P/N 分類を同時に解決する. 例えば, 以下の例で悪性腫瘍はどちらも ER で病名と判断されるが, 同時に病名の後ろに表れるコンテキストから P/N が判断される.

- *検査の結果, 悪性腫瘍が**認められた**(陽性)*
- *検査の結果, 悪性腫瘍が**疑われた**(陰性)*

表 1. 各ネットワーク層のパラメータ次元数

ネットワーク層	入力次元数	出力次元数
文字ベクトル変換	文字 ID (*1)	100
双方向 LSTM	100	100
隠れ層	200	5(*2)
CRF 層	5x 入力シーケンス長	入力シーケンス長

(*1) 文字辞書から取得された文字 ID
(*2) BIO ラベルの種類数("B-P","I-P","B-N","I-N","O")

心	電	図	で	心	房	細	動	を	認	め	た
-	-	-	-	I48	I48	I48	I48	-	-	-	-
漢	漢	漢	ひ	漢	漢	漢	漢	ひ	漢	ひ	ひ
O	O	O	O	B-P	I-P	I-P	I-P	O	O	O	O

図 3. ネットワークへの入力データと正解 BIO ラベルの例. 1行目は入力テキストであり, 2, 3 行目はそれぞれ ICD-10 情報と文字タイプ情報を示す. 最後の行は正解 BIO ラベル.

P/N 分類は, 学習データに表れる様々な陽性, 陰性の所見パターンから学習される. 逆に言えば学習データに表れない所見パターンの P/N 分類は困難となる. 学習データのアノテーションでは, 否定や不確実性を表すモダリティや直接患者と関係のない一般論を述べる文脈では陰性タグ(<N>～</N>)を付与し, 患者が現在罹患していると判断される場合は陽性タグ(<P>～</P>)を付与した. BIO ラベルでは図 1(b)に示すように陽性タグの固有表現は開始を表す“B-P”および中間を表す“I-P”, 陰性タグの固有表現は開始を表す“B-N”, および中間を表す“I-N”によってタグの範囲が示される. 固有表現に属さないそれ以外の文字は全て“O”で示される.

P/N 分類におけるモダリティの判断は, 双方向 LSTM により該当文字の前後のコンテキストに依存して処理される. 我々の事前実験では, 隠れ層の後に CRF ではなく Softmax を用いた場合, 矛盾のない BIO ラベリングが必ずしも得られないことを確認した. 例えば, B-P の後には, B-I または O が現れるべきであるが, I-N が現れる場合などである. Softmax の代わりに CRF を用いることでシーケンス全体における隣り合うラベルの遷移確率を学習することができ, これにより矛盾のない一貫した BIO 系列ラベリングを行うことが可能となる.

4. 実験

本研究では, コーパスとして NTCIR の共有タスクデータ[Morita 13]と互換性のあるアノテーション付きの 500 症例(計 5,158 文)からなる医療テキストデータセットを用いた. これらのデータには, 陽性タグ, 陰性タグが教師データとして与えられている.

4.1 評価方法

評価は, 10-分割交差検証をテストデータ入れ替えて 10 回検証を行った(最終的な評価は 10 回の平均). 評価手法は, CoNLL-2000 共有タスクと同じである. 評価に使用した Perl スクリプトは, CoNLL-2000 の Web サイト[1]から入手できる. 性能は, 既存研究の手法に基づき, 適合率, 再現率および F-尺度(β=1) を用いて評価した.

4.2 ネットワークへの入力データ

ネットワークへの入力として, 1)表層文字だけを用いる場合(Bi-LSTM), 2)事前処理により付与した ICD-10 コードの情報を追加情報として用いた場合(Bi-LSTM+ICD), 3)事前処理により付与した ICD-10 コードと文字タイプの情報を追加文字情報として用いた場合(Bi-LSTM+ICD+Ctype)3 つのケースを用いて評価を行った. 図 3 にネットワークへの入力データの例を示す. 最初の行が入力テキスト, 2 行目が ICD-10 コード情報, 3 行目が文字タイプ情報, 最後の行が正解の BIO ラベルである. 表 2 に, 入力に用いた辞書サイズとベクトルに変換後の次元を示した.

表 2. 入力データの辞書サイズと変換後のベクトル次元数

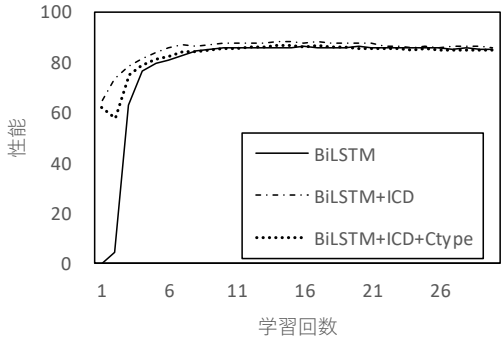
文字情報	辞書サイズ	変換後ベクトル次元数
表層文字	1516	100
ICD-10 コード(*1)	15040	100
文字タイプ(*2)	4	10

(*2) ICD10 標準病名マスター[Medis]
(*3) ひらがな, カタカナ, 漢字, 英数字

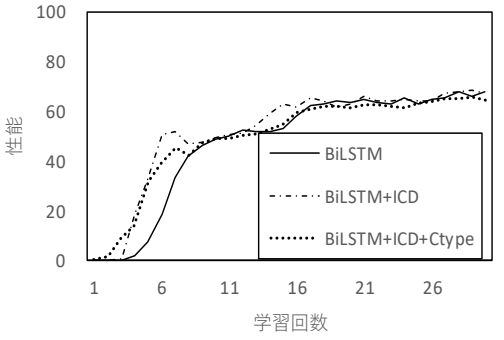
5. 結果

5.1 タグの抽出性能

図 4 にネットワークへの入力データを変えた 3 つのケース (Bi-LSTM, Bi-LSTM+ICD, Bi-LSTM+ICD+Ctype) での1から30までの学習回数における<P>タグおよび <N>タグの抽出精度 ($F(\beta=1)$)の推移を示す. 図から, <P>タグ, <N>タグの両方について表層文字以外に文字の追加情報を用いた場合, 初期の学習結果に対しては有効的に働くが, 学習回数が増えるに従って, ほとんどその効力がなくなることが分かる. このことから, 十分なコーパスがあれば, 外部リソースや知識ベースは不要な可能性がある.



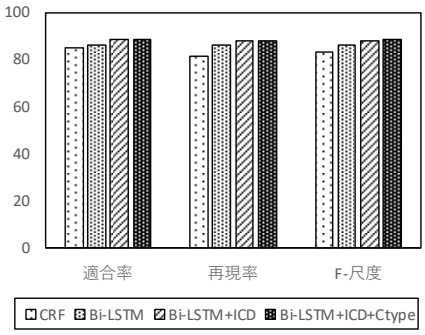
(a) <P>タグの性能($F_{\beta=1}$)



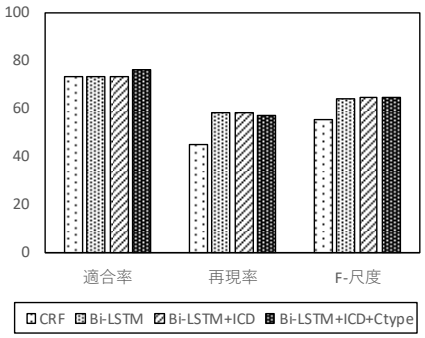
(b) <N>タグの性能($F_{\beta=1}$)

図 4. 学習回数(Epochs)に対する<P>,<N>タグ抽出の性能の変化

[1] http://www.cnts.ua.ac.be/conll2000/chunking/



(a) <P>タグ(陽性事象)の抽出性能



(b) <N>タグ(陰性事象)の抽出性能

図 5. CRF と Bi-LSTM＋CRF の性能比較. Bi-LSTM は, 入力文字のみを特徴として用いた場合で, Bi-LSTM＋ICD , Bi-LSTM+ICD+Ctype は, それぞれ追加特徴量としてそれぞれ ICD コード, ICD コードと文字タイプを用いた場合を示す.

これは, 頻度が少ない<N>タグに比べて, より頻出する<P>タグが学習回数のかなり早い時点で性能が飽和状態に達していることからも分かる.

図 5 に文字ベースの CRF[矢野 17] と, 提案手法である Bi-LSTM を用いた深層学習との性能の比較を示した. <P>タグ, <N>タグの両方について CRF よりも深層学習が高い性能を示し, $F_{\beta=1}$ 尺度で<P>タグ検出については+5, <N>タグ検出ついては+10 の精度の向上を確認した. 文字の追加情報を用いた場合, 表層文字だけを用いた場合に比べて僅かな精度向上が認められたが, その差は CRF と深層学習ほど大きいものではなかった. なお, <N>タグの精度は, <P>タグのそれと比較して低く, P/N 分類がむしろ困難な課題であることを示唆している. P/N 分類は, 人間でさえ難しい場合もあり, この結果は妥当であるといえる.

5.2 処理時間

実用的なシステムを構築するうえで, 処理時間は一つの重要な指標である. 実際に, 大学病院のような大規模な病院では毎日約 3,000 人もの患者が診療を受け, およそ 6 万もの文書が生成されている. これらの文書を処理するのにかかる時間を以下のスペックを有する計算機を用いて見積もった.

CPU: コア i7 6800K 6core/12thread 3.4GHz
メモリ:64GB (8GB×8) DDR4-2133 クワッドチャネル
GPU: GeForce GTX 1080 8GB

結果として, 医療テキスト 1,000 件の処理時間は 110 秒 (1.8 分)であった. 1 日で生成される全ての文書に対する処理時間に換算すると 109 (= 1.8 × 60,000 / 1,000) 分であった. 夜間などにバッチ処理するのであれば, 問題のない処理時間と言える. また最適化によりさらに処理時間が短縮される可能性がある. 今後の課題として, 抽出した病名, 症状名の正規化(標準化)を行う必要があり, そのための辞書も構築する必要がある.

6. おわりに

本研究では, 深層学習による (1) 事象抽出と (2) P/N 分類の 2 つの処理からなる医療テキスト解析ツールを開発した. 提案する文字ベースのアプローチは, 2 つの処理を 1 つの系列ラベリング問題として“End-to-End”で処理を行う. このアプローチには次の 3 つの重要な利点がある. 1) 形態素解析などの事前処理を必要としないため処理が簡素化された点, 2) 系列ラベリングに有効な特徴量を教師データから自動的に学習する点, また 3) コーパスサイズをさらに増やすことでさらなる精度の向上が期待できる点である.

提案手法は, 小規模なデータにおいてさえ, CRF をベースとする手法よりも高い精度を示しており, 今後, さらに大規模なデータが構築されればますます提案アプローチが有効であることを示唆している.

謝辞

この研究の一部は日本学術振興会補助金番号 JP16H06395 および 16H06399, ならびに厚生労働科学研究費補助金番号 28030301 によって支援された.

参考文献

[Chapman 11] W.W. Chapman, P.M. Nadkarni, L. Hirschman, L.W. D'Avolio, G.K. Savova, and O. Uzuner, Overcoming barriers to NLP for clinical text: The role of shared tasks and the need for additional creative solutions, *J Am Med Inform Assoc*. 18 (2011) 540-543.

[Friedman 95] C. Friedman, S.B. Johnson, B. Forman, and J. Starren, Architectural requirements for a multipurpose natural language processor in the clinical environment, *Proc Annu Symp Comput Appl Med Care* (1995), 347-351

[Ctakes] cTAKES, https://ctakes.apache.org/

[矢野 17] 矢野憲,若宮翔子,荒牧英治, 医療テキスト解析のための事実性判定と融合した病名表現認識器, *言語処理学会第 23 回年次大会*, 2017

[Chi 16] J. P.C. Chi and E. Nichols, Named Entity Recognition with Bidirectional LSTM-CNNs, *Trans. of the ACL*, vol.4, pp. 357-370, 2016

[Lample 16] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami and C. Dyer, Neural Architectures for Named Entity Recognition, *Proc. of NAACL-HLT*, pp. 260-270, 2016

[Hochreiter 97] S. Hochreiter and J. Schmidhuber, Long Short-Term Memory, Nueral Compu., vol.9, num. 8, pp. 1735-1780, 1997

[Morita 13] M. Morita, Y. Kano, T. Ohkuma, M. M., and E. Aramaki, Overview of the NTCIR-10 MedNLP task, *Proc. of NTCIR-10*, 2013.

[Medis] MEDIS, http://www.medis.jp/