Vol.2020-NL-246 No.17
Vol.2020-SLP-134 No.17
2020/12/2

診療録解析のための文のセグメント分割と意味ラベル付与

安道健一郎[1,2,a)]　奥村貴史[3,2]　小町守[1]　堀口裕正[4]　松本裕治[2]

概要：近年，医療の現場において生成される各種情報を研究活用するための情報基盤整備が進められてきた．そのなかでも，電子カルテに含まれる各種テキスト情報は，患者情報の中心であるため効率的かつ高精度な解析が必要である．そのため，多様な医療用自然言語処理研究が進められているが，多くの研究は主に文単位で行われてきた．しかし，文が医療文書の処理に際しての効率的な処理単位であるかは必ずしも自明ではない．
そこで本研究では，効率的な電子カルテの処理に向けて，記載内容を意味単位で捉えるために文をさらに分割するセグメントを定義し，その単位での意味的解析に試行的に取り組んだ．まず，医療文書で用いられる意味的な最小単位に合致するセグメントを定義し，ダミーカルテを対象にセグメント境界をアノテーションした．その上で，各セグメントに対して内容に即した 10 種類の意味ラベルを付与した．さらに，それらを自動分類するようなモデルを構築し，セグメント分割では F1 値 0.85，意味ラベル付与では平均 F1 値 0.66 で分類されることを確認した．

1. はじめに

医療の情報化に向けた各種施策により，近年，医療の現場において生成される情報が電子化され，各種の研究に活用するための基盤が整備されつつある[*1] [*2]．こうした情報基盤には，画像データや検査結果，属性データなど多様なデータが含まれ，その解析に基づく研究が積極的に進められている [1][2][3][4]．しかしながら，集積される情報のうち，電子カルテに含まれる各種の自由記載文を対象とした研究は比較的少ない [5][6][7]．

電子カルテには，患者に関する詳細かつ多様な情報が含まれる．その中でも，自由記載文はカルテの本体とも言うべき患者情報の主体であり，貴重な資源と言える．それにもかかわらず，研究利用が進まない原因として，構造化されていない自由記述部分は扱いが難しく，既存の自然言語処理技術での高精度な解析が困難であることが挙げられる．そこで，医療用自然言語処理研究として，ダミー文書を利用して記述内容の構造化を行う研究や，記述内容から目的の情報を抽出する研究が行われてきた [5][7]．

これらの研究においては，入力テキストとして一般に文が使われてきた．しかし，医療文書の特性として，検査結果，診断，今後の方針など，性質が異なる内容の記述が一文に混在している点が挙げられる (図 1・最上部)．また，実際の診療録は，図 2 に示されるように，検査結果や治療経過の箇条書き等，一般的な文章とはその言語的な性質が大きく異なる．そのため，医療用自然言語処理における処

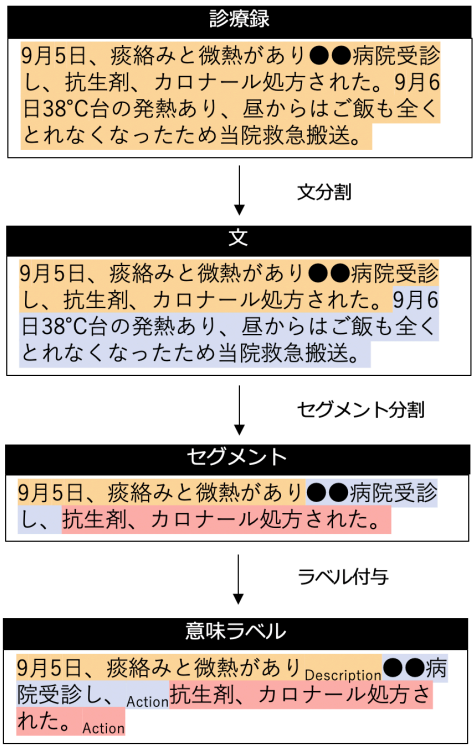

図 1: 本研究の概観

1 東京都立大学 / Tokyo Metropolitan University
2 理化学研究所 革新知能統合研究センター / RIKEN AIP
3 北見工業大学/ Kitami Institute of Technology
4 国立病院機構 / National Hospital Organization
a) ando-kenichiro@ed.tmu.ac.jp

*1 診療情報集積基盤，https://nho.hosp.go.jp/cnt1-1_000070.html [accessed 2020-10-8]
*2 診療録直結型全国糖尿病データベース事業，http://jdreams.jp [accessed 2020-10-8]

理対象として，文が適切な単位であるかは自明ではない．むしろ，適切な細分化により，既存の文章と特性が大きく異なる医療文書をより正確に解釈できる可能性がある．これは関連するタスクにとって広く有用である．

そこで本研究では，文書に何が記述されているかを分析するために，医療テキストに記述内容をどのような粒度で扱うべきかを検討した．そのために，図 1 に示すパイプラインの検証を試みた．まず，最初に診療録文書を文単位に分割する処理を行う．次に，文をさらに細かくするような**セグメント**に分割する．最後にそのセグメントに対して意味を表すラベルを付与し，文に内包されている意味セグメントを明らかにする．意味セグメントは内容に即した**意味ラベル**と，記述者の主観によって定義される**事実度**を持つものとした．

以下では，まず，2 章において解析対象とする文書について改めて定義する．3 章において，セグメント分割ルールについて詳述し，4 章において，定義した意味ラベル及び事実度を概説する．5 章において，自動分類のために構築したモデルとその検証結果を示す．6 章にまとめと展望を記す．

本研究には，以下のような貢献がある．

- 文を予め想定した意味単位で分割できるようなセグメントを定義し，ダミーカルテに対してセグメント境界のアノテーションを行った．そのデータをテストデータとして複数のモデルを比較し，アノテーションを学習データとして用いた RNN によるモデルを使用して高い精度で分割できることを確認した．
- 医療文書における最小の意味単位を想定した意味ラベルを定義し，セグメントに対して意味ラベルのアノテーションを行った．そのデータを学習データとして BERT モデルを用いた分類モデルを作成した．
- ダミーカルテにアノテーションされた意味ラベル分布の分析を通して，医療文書中にどのような記述がどの程度含まれているかを示唆した．

2. 本研究での解析対象

解析対象文書

解析対象とする文書について説明する．本研究では，診療録とは医師が診療の際に作成する文書と定義する．そのような医師が診療時に作成する文書として主に，(1) 外来カルテ (2) 入院カルテ (3) 退院時サマリの 3 つが存在する．

(1) の外来カルテとは，患者が外来で医療機関を受診した時に作成される文書である．主訴や既往歴，現病歴，身体所見，検査所見，治療計画などが記述され，患者の治療記録として保管される．(2) の入院カルテは患者が入院した際に作成される．記述項目は外来カルテと本質的には等価

#1 細菌性髄膜炎
4/20〜5/8　VCM 1250mg(q12h)
4/20 SBT/ABPC 1.5g単回
4/20〜 MEPM 2g(q8h)
4/20〜4/23 ﾃﾞｷｻｰﾄ 6.6mg(q6h)
4/20〜4/22 日赤ﾎﾞﾘｸﾞﾛﾋﾞﾝ
4/20 腰椎穿刺1回目　髄液 糖定量 30 mg/dl(血中糖 95mg/dl) 細胞数 2475/μl.
グラム染色するも明らかな菌が見つからず、髄液培養でも優位な菌は培養されなかった。
細菌性髄膜炎に対するグラム染色の感度は60%程度であり、培養に関しても感度は高くない。
また髄液中の糖はもう少し減るのではないだろうか。
確定診断はつかないものの、最も疑わしい疾患であった。
起因菌はMRSA,腸内細菌等を広域にカバーするためバンコマイシン,メロペネム(髄膜炎dose)とした。

図 2: 診療録の例

であるが，検査や処置，日々の回診毎に記載されていくことで密度が大きく異なる．そのため，入院経過が長くなるにつれ，記述量が膨大となりうる．(3) の退院時サマリは入院カルテを要約したものである．診療録の用途として，診療行為に記録を残すというものがあるが、入院等では重要な検査結果等に加えて，日々の記録など重要でないさまざまな情報が生じうる．退院後，そのように記述量の多い入院カルテを参照するのは非効率であることに加えて，紹介での入院等においては紹介元の医師に入院経過を効率的に知らせる必要がある．そこで，退院毎に，医師の手により生成することが義務付けられているのが退院時サマリである．

本研究では研究で用いられることが多く，比較的記述の乱れが少ない (3) の退院時サマリを解析対象の文書とする．

解析対象セクション

解析対象とするセクションについて説明する．退院時サマリは電子カルテを使用して構造化して書かれることが多い．その際，事前に用意しておいた電子カルテシステムのフィールドから各項目に記述が流し込まれて初期文書が生成されることが多い．その例を図 3 に示す．

システムより流し込まれて自動生成された記述には意味のまとまりがある項目とまとまりのない項目が存在する．意味のまとまりがある項目は身体所見や検査所見などが顕著である．一方，現病歴，入院中経過記述などは身体所見や検査所見，診断などが混在しており，意味的まとまりのないセクションである．本研究では意味が異なる記述が混在した文を意味単位で解析するという目的があるため，以降では現病歴と入院中経過記述セクションに属する文章を対象テキストとして扱う．

3. セグメント分割

診療録のきめ細かい解析を実現するためには文より細かい粒度で記述内容の意味を捉えることが必要である．そのため，この章では診療録の文章を文に分割し，その文をさ

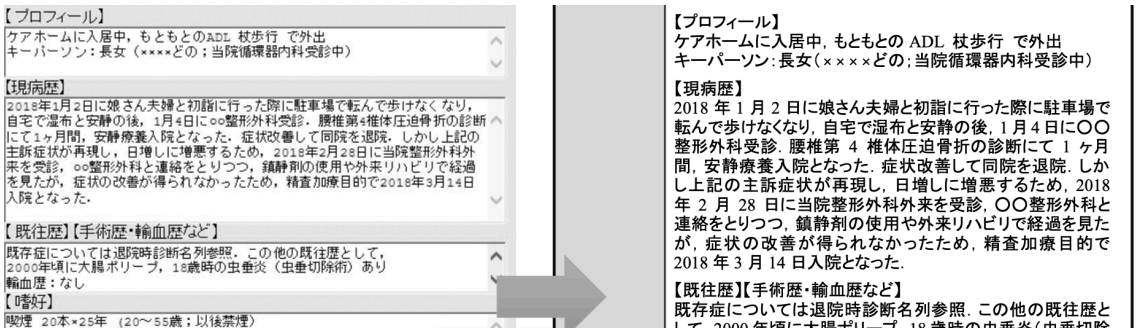

図 3: 電子カルテシステムを用いた退院時サマリ生成の例 [8]

らに細かい単位のセグメントへと分割する方法を説明する．はじめに，文章から文の分割について述べる．

文分割

医師が記入する診療録は図 2 に一部確認できるように，句点が抜けていたり，読点で改行されていたり，空白が句読点の代わりをしていたり，ノイズが多い文章である．そのため，どういう記述を文として扱うかを定義する必要がある．よって，文分割の境界として次に挙げる 2 つのルールを定義した．(1) 「。」や「.」の句点で終わる．(2) 改行で終わり，「、」や「,」の読点が付かない．2 章で挙げた対象セクション内の文章にこれらのルールを適用したものを解析対象の文とする．

セグメント分割

セグメントの定義を説明する．基本的には 4 章で説明する意味ラベルを付与できるような単位を最小単位として定義し，分割する方針とした．次に分割ルールの詳細を示す．例中の ⟨SEP⟩ はセグメントの境界を表す．

(1) 述語，読点，事態性のある名詞句で分割
セグメント分割におけるベースのルール．このルールを基準とし，他のルールを例外として適用する方針とする．
（例：絶食、⟨SEP⟩ 抗菌薬投与で ⟨SEP⟩ 肺炎は軽快。）

(2) 括弧の中に文が内包されている場合は括弧の記号で分割する
括弧の中の (1) で分割されるような表現を取り出すため，括弧内に分割されうる対象が存在する場合は括弧の記号をセグメント境界とする．
（例：画像で「⟨SEP⟩ 両側肺門部に陰影あり、⟨SEP⟩CT で両肺に多彩な浸潤影を認め ⟨SEP⟩ 重症肺炎」⟨SEP⟩ として 4 月 10 日に入院。）

(3) 疾患名は積極的に分割する
疾患名は診断として書かれることが多く，(1) のルールで分割するところでなくても重要な意味を持つ．よって疾患名を含む記述はセグメントとして分割する．
（例：肺炎疑いで ⟨SEP⟩ 当院紹介となった。）

(4) 検査結果とその評価は分割する
診療録中には検査結果とその評価が書かれることが頻繁にある．検査結果の記載と結果に対しての評価は明確に意味が異なるため，(1) に当てはまらなくても分割する．
（例：血清クレアチニンキナーゼは 4512 U/L と ⟨SEP⟩ 高度に上昇していた。）

(5) 治療に関係ないものは分割しない
医療的に意味がある事象を分割の対象としているので，医療行為と関わりの薄い描写は (1) に該当しても分割しない．
（例：ケアマネジャーに同伴されて来院した。）

(6) 意味の増加がない表現は分割しない
「〜の方針」,「〜治療を継続」など，前の記述を受けて意味を補足するような表現は分割しても意味の増加がないと考え，(1) に該当しても分割しない方針とした．
（例：外来で抜糸を行う方針とした。）

(7) 主に名詞句が列挙された際，意味ラベルが変化しないような記述は分割しない
意味的に同じような記述，つまり意味ラベルが変化しないような記述が列挙されている場合は，分割して個別にする意義があまりないので，分割しない．具体的には，検査の列挙や症状名の列挙などが該当する．しかし，(1) の例のように，意味ラベルが異なるような記述は分割する．
（例：発熱、盗汗、体重減少、喀痰、血痰は否定。）

(8) 時間関係を表す記述は分割しない
時間の経過などの表現は前述の内容を受けて書かれることが多い，または仮定の事象を表すことが多いので，分割対象としない．
（例：抗菌薬開始後、発熱・腹痛は徐々に改善し）
（例：入院後、CMZ で抗菌薬加療を開始し、）

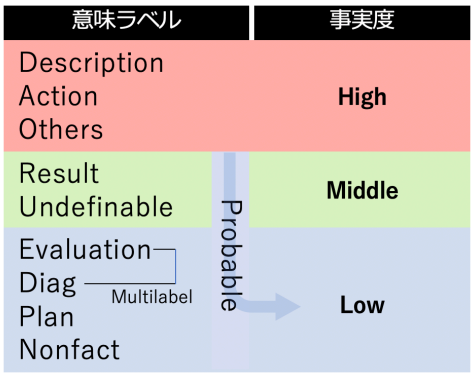

図 4: 意味ラベル構造

4. 記載内容の意味分類

4.1 意味ラベル

次に，上述したセグメントそれぞれについての意味分類として定義した 10 種類のラベルを説明する．

ラベルの外観を図 4 に示す．

Description

過去の出来事を描写する記述．

診療録に出現する基本的な記述である．見たものをそのまま描写していて，患者の身体所見や，治療中の出来事の記載，検査結果の羅列，検査結果の言い換え（高い，低い，陽性，陰性など），患者の過去のエピソードも含まれる．

後述する Action のような，なにかの行為を表すような記述でも，受動態でなおかつ行動主が医療者である場合は患者の経験と捉え，Description ラベルを付与する．

現病歴の記載内容は患者の過去のエピソードがほとんどを占めるため，この過去の事実を描写する記述が多い．

（例：尿中肺炎球菌抗原陽性のみ出ている。）

Action

過去に何者かが行動したことを描写する記述．

「入院した」や「投与する」などが代表的な例である．行動主が患者，患者以外の場合があり，「（行為）のようだった」のような行動を出来事の描写として記述している，または「（行為）された」のような受動態の場合は行動主が患者に限り，Action ラベルを付与する．行動主が患者以外の場合は Description とした．

（例：経過良好で退院した。）

Others

セグメント分割上出現した意味を持たないセグメント．括弧で分割した場合などが含まれる．

（例：しかし「）

Result

基本的には事実の描写であるが，医療者の主観が入っていると言えるもの．

たとえば，CT 検査等の画像診断においては，基本的に画像に示されている異常を所見としてそのまま記述することになる．しかしながら，その多くは定性的な表現を含むことから，血液検査のように定量的に結果が示される検査と異なり，その評価に観察者の主観が入り込む余地がある．そのため，後述する「事実度」の分類を行うのに際して，事実度の高い客観的な記述と推論等の事実度の低い記述の中間的なカテゴリを設ける必要があった．

また，検査結果の「高い」「低い」や「陰性」「陽性」などは事実の描写と捉えられるが，数日にわたる数値の変化を「改善」「悪化」と描写する際も，事実ではあるものの観察者の主観が混入する余地がある．前医が下した診断結果に関する記載なども，客観的な診断結果か否かの判定が難しいためケースがありうる．そこで，そのような記述に対して Result ラベルを付与するルールを設けた．

（例：右下肺野に浸潤影あり、）

Undefinable

基本的には事実の描写であるが，将来への言及と受け取れるもの．

「～退院.」や「～となった」などは表現だけでは将来への言及か過去の事実の描写か判断できない．文脈を考慮すれば判断できる例もあるが，アノテーターによって揺れるため Undefinable のラベルとする．事態性名詞の体言止めが多く含まれる．

基本的には High，Low のボーダーケースで，表現が省略されたため将来に対する記述か判定できなくなったものが属する．

（例：として 4 月 10 日に入院。）

Evaluation

検査結果などに対して明らかに医療者の主観や推論が入っている記述．

診療録の核となる記述で，しばしば臨床診断と合わせて記されるため，後述する「Diag」とマルチラベル構造をとる．頻出する例としては，検査結果が羅列された後にそこから得られる所見を述べたもので，検査結果から推論を経て所見を生成したような流れで記述される．

（例：間質性肺炎の急性増悪と考えられたことから）

Diag

臨床診断を下している記述．

これも Evaluation と同じく診療録の核となる記述であり，Evaluation ラベルとマルチラベル構造をとる．基本的には臨床診断だけで分割されたようなセグメントであるが，セグメント分割の結果，所見と臨床診断を分割できない例がある．その場合は所見から推論を通じて臨床診断を得たようなセグメント構造になる．

「呼吸困難」のように症状と疾患名のどちらも該当するものが存在する場合はボーダーケースと捉え，Result のラベルを付与している．

(例：臨床的に肺小細胞癌 ED と診断された。)

Plan

明確に将来についての言及をしている記述.

「方針」「予定」などの明確に将来の予定について記述しているセグメントが属する．主に入院経過記述の後半に記載され，次回治療の予定や退院後の患者の所在の予定について書かれることが多い．医療従事者が読めば退院後について書かれていると分かる記述も多いが，アノテーションの揺れを抑えるため表現だけでわからないものは Undefinable ラベルを付与することとしている.

（例：待期的 PCI の方針となった。）

Nonfact

伝聞表現や仮定などで書かれ，事実をもとにした記述でないものが含まれる.

Evaluation，Diag，Plan のどれにも属さないが，言語表現で明らかに事実でないことが分かるセグメントである.

（例：病理解剖を打診したかった症例ではあるが、）

Probable

言語表現で事実でない，主観が主であることがうかがえるような記述.

「疑い」や「思う」などが典型例で，後述の事実度 Middle，Low に補足的に付加される．よって，Result，Undefinable，Evaluation，Diag，Plan，Nonfact ラベルとマルチラベル構造を取る．Probable ラベルが付与されたセグメントは後述する事実度 Low に分類される.

（例：当初は腎膿瘍または腎細胞癌を疑った。）

4.2 事実度

以上の 10 種類のラベルを，記述の確からしさに応じて，さらに 3 段階の事実度に分類した.

High

Description，Action，Others が属する．事実を見たまま描写したものであることが明確，記述者の主観が入っていないことが明確なセグメントである．診療録の記載の基本を構成するセグメントであることが予想される.

Middle

Result，Undefinable が属する．High にも Low とも断定できないセグメント．主にアノテーションの揺れを抑える目的で設けられたラベルで，後の解析フェーズでのノイズを減少させる意図がある.

Low

Evaluation，Diag，Plan，Nonfact，Probable が属する．明らかに事実の描写ではなったり，記述者の主観を中心に置いた記述が該当する.

所見などが多く含まれ，事実に対しての推論や外部知識の導入などが頻発しているセグメントであることが予想される．入院経過記述の核をなす記述である.

表 1: 対象診療録の統計情報

症例数	文数	文字/文	高頻度疾患（症例数）
108	1,748	35.58	誤嚥性肺炎 (10), 尿路感染症 (6), 気管支肺炎 (3), 急性心筋梗塞 (3), 虚血性腸炎 (3), 細菌性肺炎 (3)

5. 実験

前述の定義に従ってセグメントと意味ラベルについてアノテーションを行い，分類モデルを作成した.

実験で共通して用いたダミーカルテ [9] の統計量を表 1 に示す．ダミーカルテは医師によって架空の患者を想定して作成された．ダミーカルテは全 108 症例で，自由記述部分 1,748 文で形成される．最も出現する疾患は誤嚥性肺炎であり，全て内科で作成された文書である．自由記述部分に当たる現病歴と入院経過記述は医療従事者の手によって退院時サマリより抽出された.

5.1 セグメント分割

5.1.1 アノテーション

3 章で述べたセグメントの定義に則り，文に対してセグメント境界のアノテーションを行った．まず著者 1 名がアノテーションを行い，ラベル一致率確認のために追加で医療従事者 1 名の協力を得た．結果，アノテーション境界における著者と医療従事者間の一致率は 0.82 となった．ラベルは高い一致率でアノテーションされており，客観的なラベルであることが確認された．また，総セグメント数は 3,816，1 文あたりの平均セグメント数は 2.18，1 文あたりの平均セグメント境界数は 1.18 であった.

自動分割モデルのためのデータセットは文をシャッフルし，訓練：開発：テスト ＝ 1,548：100：100 としたものを使用した.

5.1.2 分割モデル

分割モデルのベースラインとして，読点のみで分割したモデル（Punct），読点および動詞[*3]で分割するモデル（Punct&Verb）を用いる．また，セグメントの定義は文を分割するという点において「節」に類似していると思われるため，節をルールベースで分割するモデル（CBAP [10]）も採用した．このモデルは形態素情報を基にした 332 個のルールで構成されている.

機械学習モデルとして，談話構造解析において談話ユニットを分割するために開発された SEGBOT [11] を用いた．このモデルは，分割対象の N 個のトークンを入力とするシーケンスを $\mathbf{U} = (U_0, U_1, ..., U_N)$ と定式化すると，U_0，U_4，U_7 がセグメント開始トークンの場合は図 5 のよう

[*3] 動詞を起点として，次に出現する非自立を除く名詞の前で分割する.

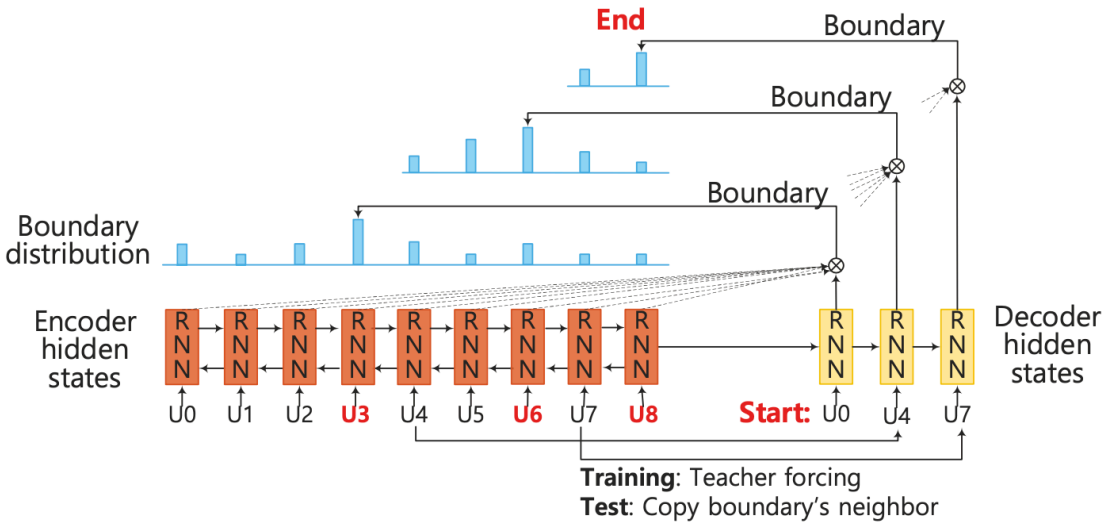

図 5: RNN 分割モデル [11]

表 2: セグメント分割の実験結果

モデル	Precision	Recall	F1
Punct	0.521	0.187	0.275
Punct&Verb	0.569	0.610	0.589
CBAP [10]	0.280	0.136	0.183
SEGBOT [11]	**0.864**	**0.829**	**0.846**

に表せる．入力シーケンスは MeCab[*4]で形態素に分解し，辞書として mecab-ipadic-NEologd [12] を用いた．形態素を Wikipedia と Web ページで学習済みの FastText [13] によって 300 次元のベクトルに変換した．また，SEGBOT の訓練の際，埋め込み層は追加学習せずに固定した．

5.1.3 結果

実験結果を表 2 に示す．RNN ベースの SEGBOT が一貫して高精度であり，次いで精度が良い読点と動詞で分割したモデルより F1 値が 0.26 ポイントほど高い．この結果より，セグメント分割でも RNN を用いることが有用であることが示された．また，節分割モデルである CBAP が F1 値 0.183 と低いことから節の定義と本研究におけるセグメントの定義は本質的に異なるということが示された．読点で分割したモデルの Precision が 0.5 ほどしかないことから，本研究のセグメントは読点で必ず分割されるわけではなく，前後の文脈によって分割されうるか決定されていることが示されている．

また，実験のテストセットにおけるエラーの例を表 3 に示す．表中には ⟨SEP⟩ を用いてモデルが予測した境界，正解の境界，それが意味分割において許容できるか否かを医療従事者と共に判定した結果が記してある．結果，テストセット 100 文に含まれる予測境界のうち，許容できないような分割は 2 箇所のみだった．この 2 箇所以外は粒度が粗くなったり，逆に細かくなったりするものの，セグメントに対して何かしらの意味ラベルを付与できるような分割境界になっていた．許容できない例の 1 つである最上部の例は疾患名をうまく捉えられていないことに起因したエラーである．医療用語を補強するような辞書を MeCab に用いて形態素に分解すればこのようなエラーは生じないと考えられるため，今後の課題とした．最上部から 3 段目の例は細分化しすぎた例である．最下部の例は前処理で分割すべ空白を除去している影響で境界を予測できなかった例である．今回のエラーで最も多かったのがこの空白の除去で分割できなかった例であった．そのため，空白を 1 つのトークンとして処理し，SEGBOT を再訓練する追加実験を行った．結果は F1 値が 0.688 と空白を考慮しないモデルに比べて 0.16 ほど低下した．理由として日付や検査名と値の間に空白が入れられることが多く（表 3 中の 1/4，Hb と 8.2），それらがノイズとなっていることが考えられる．

5.2 意味ラベル付与

5.2.1 アノテーション

セグメント分割と同じように 4 章の定義に則り，セグメントに意味ラベルを付与した．アノテーションは医療従事者 2 名により行った．アノテーションされたラベルの分布を表 4 に示す．ラベルは過去に起きた事象の描写を表す Description が一番多く，次に Action であった．いずれも過去の事実に関する記録であり，診療録の本来の目的からも妥当であると考えられる．また，Description よりも Action が少ないことから，診療録に何者かの行動に関する記述よりも既成事実の記載が 2 倍ほど存在することが示唆される．

*4 https://taku910.github.io/mecab

表 3: セグメント分割のエラー

予測境界	正解境界	判定
2014 年 4 月に発症した間 ⟨SEP⟩ 質性肺炎のため ⟨SEP⟩	2014 年 4 月に発症した間質性肺炎のため ⟨SEP⟩	×
痙攣に関与するよう ⟨SEP⟩ な部位でも大きさでもなかった。	痙攣に関与するような部位でも大きさでもなかった。	×
先行する感冒の ⟨SEP⟩ 病歴と ⟨SEP⟩ 肉眼的血尿に伴って ⟨SEP⟩	先行する感冒の病歴と肉眼的血尿に伴って ⟨SEP⟩	○
1/4Hb8.2EGD：胃潰瘍 (A1stage：体上部後壁)	1/4Hb8.2⟨SEP⟩EGD：胃潰瘍 (A1stage：体上部後壁)	○

表 4: ラベルの分布

事実度	意味ラベル	数 (%)	
High	Description	1,465(37%)	2,328(61%)
	Action	798(20%)	
	Others	65(2%)	
Middle	Result	306(8%)	642(17%)
	Undefinable	341(9%)	
Low	Evaluation	278(7%)	846(22%)
	Diag	255(6%)	
	Plan	265(7%)	
	Nonfact	82(2%)	
	Probable	134(3%)	

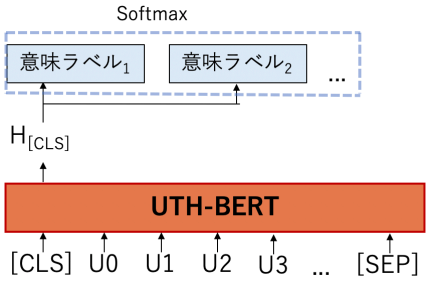

図 6: 意味ラベル付与モデル

事実度低の中では Evaluation，Diag，Plan がほとんど同じ数であり，Nonfact や Probable は相対的に少なかった．これは，診療録中に所見や検査結果に対する評価や臨床診断，将来の治療予定に関する記述が等しい量含まれていることを示唆するものである．

ラベル付与モデルに用いるデータセットとしてセグメントをシャッフルし，訓練：開発：テスト ＝ 3,616：100：100 を用いる．

5.2.2 分類モデル

分類モデルとして診療録で事前学習された BERT [14] である，UTH-BERT [15] に分類層を追加したものを使用した．モデルの概要を図 6 に示す．UTH-BERT は BERT における $\text{BERT}_{\text{BASE}}$ と同サイズである．UTH-BERT の学習設定に従い，セグメントの単語分割には MeCab を用いて，追加の単語辞書として mecab-ipadic-NEologd [12] と J-MeDic [16] (MANBYO_201905) を用いた．

また，BERT を意味ラベルで追加学習した．セグメントを BERT への入力とし，意味ラベルの分類層に入力するベクトルとして BERT の入力トークンにおける [CLS] に相当する最終隠れ層を用いて行った．

意味ラベルは 4 章で述べたとおり，Probable が，Low ラベル，Middle ラベルへのマルチラベルになる構造になっており，全ての意味ラベルについて上位階層に事実度のラベルが存在する．そのため，ラベル付与タスクを次のように 3 つに分割した．(1) Probable 以外の 9 ラベルを対象にした 9 値分類（**意味ラベル分類**）．(2) Probable に該当するかどうかの 2 値分類（**Probable ラベル分類**）．(3) 事実度を対象にした 3 値分類（**事実度分類**）．

意味ラベル分類については図 6 の構造を用い，Diag と Evaluation が同時に付与されている例については診断についてのラベルである Diag が主な内容であると判断し，Diag に統合した．タスクはラベルを $y \in \{\text{Description, Action, Others, Result, Undefinable, Evaluation, Diag, Plan, Nonfact}\}$ とし，入力セグメントを x としたときに，セグメントがラベルに属する確率 $P_{label}(y|x)$ を求めるものである．これらは以下のように表される．

$$P_{label}(y|x) = \text{softmax}(Wh_{[CLS]}) \quad (1)$$

$$\hat{y} = \arg\max_{y} P_{label}(y|x) \quad (2)$$

$W \in \mathbb{R}^{c \times d}$ は分類層のパラメータであり，c は分類ラベルの候補数，d は UTH-BERT 最終層の次元数である．Probable ラベル分類についてはラベルを $y \in \{\text{positive, negative}\}$ の 2 値分類に変更したものである．事実度分類についてはラベルを $y \in \{\text{High, Middle, Low}\}$ に変更したものである．

5.2.3 結果

各タスクの結果を表 5 に示す．なお，平均 F1 値は各ラベルの F1 値の加重平均である．加重平均 F_{ave} はラベル集合を y，ラベルに属する例の個数を N，ラベル内の F1 値を S とすると以下のように表せる．

$$F_{ave} = \frac{\sum_{S,N \in y} SN}{\sum_{N \in y} N} \quad (3)$$

意味ラベル分類については，Description が高い精度を示した．しかし，平均 F1 値が 0.66 と全体的に物足りない精度であり，今後さらなる精度の向上が必要であると考えら

表 5: 意味ラベル付与の実験結果

意味ラベル分類

ラベル	Precision	Recall	F1
Description	0.76	0.81	0.78
Action	0.61	0.58	0.59
Others	0.00	0.00	0.00
Result	0.75	0.33	0.46
Undefinable	0.75	0.60	0.67
Evaluation	0.62	0.62	0.62
Diag	0.33	0.67	0.44
Plan	0.56	1.00	0.71
Nonfact	0.00	0.00	0.00
平均			0.66

Probable ラベル分類

Probable	Precision	Recall	F1
Positive	0.67	0.67	0.67
Negative	0.99	0.99	0.99
平均			0.98

事実度分類

事実度	Precision	Recall	F1
High	0.80	0.74	0.77
Middle	0.33	0.56	0.42
Low	0.74	0.70	0.72
平均			0.72

れる．Result の高い Precision に関しては，Result に属するセグメントに画像に対しての評価を与える記述が多いことが起因しており，画像所見に対する表現をうまく学習できていることがうかがえる．また，Plan の高い Recall は「方針」や「予定」など特定のキーワードに対して強く学習していることが影響していると考えられる．

Probable ラベル分類については Negative について高い精度で分類できており，Positive については精度が低い．しかし，Probable ラベルは Negative：Positive = 3,683：134 の不均衡ラベルのため，これ以上の Positive の精度改善については学習方法を変更するなどの工夫が必要である．また，表 6 に示すように可能性表現の否定をうまく捉えられないなどの傾向が伺えた．

事実度分類は High と Low の精度が概ね高く，Middle が低い．しかしながら，Middle は混乱を避けるために設けたラベルでもあるので，解析の際は High と Low のみを考慮することが可能である．そのため今回のラベル付与タスクの中では最も良い結果であると言える．

6. まとめと展望

本研究ではより細かい診療録分析のため，文をさらに分割するようなセグメントの付与とそのセグメントの意味を表すようなラベルを付与し，ラベルを自動で分類するようなモデルを構築した．実験の結果，セグメント分割のアノテーションは十分な一致率であり，自動分割において RNN で良い精度を達成した．意味ラベル付与について，UTH-BERT を用いた分類を行った．事実度分類については良い精度を達成したが，他のタスクについてはラベルによっては F1 値 0.5 を下回ったりと，精度向上の余地を残す結果となった．また，診療録中に出現する記述について，既成事実の描写が最も多く，何者かの行動についての記述が次いで多いことがわかった．診療録は事実を中心に，その事実から推論などを通じて診断や治療方針が記述されることが示唆される結果となった．

今後の展望として，実際の医療文書を用いた分析を予定している．今回の小規模ダミーカルテを用いて示唆された記述分布が実際のより大規模診療録でどのようになっているのかを確認したい．また，書かれた機関や医師などによって記述分布が変わると思われるので，加えて解析したい．

より大きな診療録データセット分析のために，分類モデルの精度向上も今後の展望として挙げられる．今回の意味ラベル分類精度ではエラー量が多く，分析に影響が出ることが予想されるので，モデルの改善を通して精度向上を図りたい．一つの例に複数のラベルが付与されるので，マルチタスク学習や文脈を考慮した学習などを考えている．また，セグメント分割モデルでは，疾患名を認識できないことによるエラーが見られたため，医療用語を考慮したモデルの学習が有効であると思われる．加えて，精度低下の原因の一つに，日本語の医療用語を含んだコーパスで事前学習済みの単語ベクトルについて，公開されているものが存在しなかったために単語ベクトルの質がよくないことも考えられる．この点については UTH-BERT を用いた手法を考えたい．

また，実際にパイプライン処理としてセグメント分割を行った後に意味ラベル付与を行った際は現在の意味ラベル精度がさらに低下することが予想されるので，合わせて実験を行いたい．

謝辞

ダミーカルテ整理とアノテーション作業において，田鎖麻衣さん，中込暢子さんに大変ご尽力いただきましたことを深く感謝いたします．

参考文献

[1] 野里博和，近藤堅司，河内祐太，坂無英徳，村川正宏，小澤　順，清野正樹，藤本真一，田中雅人，安達登志樹，伊藤春海，木村浩彦：胸部 X 線右肺底領域における肺血管正常モデルに基づく病変検出，2019 年度人工知能学会全国大会（第 33 回） (2019).

表 6: Probable 分類のエラー

セグメント	予測ラベル	正解ラベル
尿路感染については施設入所者であり、緑膿菌感染の可能性が否定できなかったため、	Positive	Negative

[2] 八重樫文絵，荒木雅弘，岡　夏樹，新谷元司，吉川昌孝：レセプトデータを用いた生活習慣病の発症予測，2019 年度人工知能学会全国大会（第 33 回） (2019).
[3] 村田知佐恵，山下和也，阪本雄一郎，櫻井瑛一，本村陽一：重み付き PLSA を用いた敗血症患者の DPC データ分析結果における各クラスタの特徴およびクラスタ遷移パターンの検討，2019 年度人工知能学会全国大会（第 33 回） (2019).
[4] 小須田玲花，小名木佑来，太田丞二，高橋　愛，髙岡浩之，横田　元，堀越琢郎，森康久仁，須鎗弘樹：LSTM を用いた機能的な冠動脈有意狭窄の分類，2020 年度人工知能学会全国大会（第 34 回） (2020).
[5] Aramaki, E., Miura, Y., Tonoike, M., Ohkuma, T., Masuichi, H. and Ohe, K.: TEXT2TABLE: Medical Text Summarization System Based on Named Entity Recognition and Modality Identification, *Proceedings of the BioNLP 2009 Workshop* (2009).
[6] Sakka, K., Nakayama, K., Kimura, N., Inoue, T., Iwasawa, Y., Yamaguchi, R., Kawazoe, Y., Ohe, K. and Matsuo, Y.: Character-level Japanese Text Generation with Attention Mechanism for Chest Radiography Diagnosis, *AAAI Workshop on Artificial Intelligence* (2020).
[7] 柴田大作，若宮翔子，伊藤　薫，荒川　豊，吉江智秀，荒牧英治：電子カルテテキストを自動臨床データベース化する要約システムの開発，日本医療情報学会「医用知能情報学研究会」人工知能学会「医用人工知能研究会」(SIG-AIMED) 合同研究会 (2018).
[8] 退院時要約等の診療記録に関する標準化推進合同委員会：退院サマリー作成に関するガイダンス (2019).
[9] 安道健一郎，奥村貴史，小町　守，松本裕治：確信度に基づく退院時サマリの分析，情報処理学会第 240 回自然言語処理研究会 (2019).
[10] 丸山岳彦，柏岡秀紀，熊野　正，田中英輝：日本語節境界検出プログラム CBAP の開発と評価，自然言語処理 (2004).
[11] Li, J., Sun, A. and Joty, S.: SegBot: A Generic Neural Text Segmentation Model with Pointer Network, *Proceedings of the 27th International Joint Conference on Artificial Intelligence and the 23rd European Conference on Artificial Intelligence* (2018).
[12] 佐藤敏紀，橋本泰一，奥村　学：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 (2017).
[13] Grave, E., Bojanowski, P., Gupta, P., Joulin, A. and Mikolov, T.: Learning Word Vectors for 157 Languages, *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)* (2018).
[14] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (2019).
[15] Kawazoe, Y., Shibata, D., Shinohara, E., Aramaki, E. and Ohe, K.: A Clinical Specific BERT Developed with Huge Size of Japanese Clinical Narrative, *medRxiv* (2020).
[16] Ito, K., Nagai, H., Okahisa, T., Wakamiya, S., Iwao, T. and Aramaki, E.: J-MeDic: A Japanese Disease Name Dictionary based on Real Clinical Usage, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (2018).