1R−7

医療論文抄録からの情報抽出のための並列表現分析 ～診断，治療の評価項目を含む表現～

井上 大悟* 永井 秀利* 中村 貞吾* 野村浩郷* 中島 律子† 大貝 晴俊‡

九州工業大学 情報工学部* 科学技術振興事業団† 新日鉄‡

1 背景

近年，医療分野における診断機器や治療方法の向上は著しく，それにコンピュータ，ネットワークの発展，普及も重なって，カルテや医療文献など多種多様な医療情報が電子化されるようになった．このため多くの電子化された文書から計算機によって情報を抽出するシステムが実現することによって得られるメリットは非常に大きいため，現在盛んに研究されている．

これまでの研究で，抽出したい情報項目が1文中に1回しか出現しないような場合，それらの文には定型性が見受けられ，抽出する項目とその周辺の文字列を記述した“テンプレート”を使用することで，ある程度の抽出結果を得ることができた [1]．しかし，医療論文抄録は限られた字数で記述するため，診断，治療を行った結果部分について記述している一文においては抽出したい情報が詰めこまれていることが多い．そのため，抽出したい情報項目が1文中に2回以上出現する場合，先述した手法ではよい抽出結果を得ることができなかった．そこで，本稿では設定した抽出項目の並列表現に対し抽出を行うことを目的としている．

2 医療論文抄録と抽出項目の設定

今回対象とした医療論文抄録は1999年度日本医学放射線学会学術発表会抄録236稿と2000年度分573稿の合計809稿である．

また，抽出の対象とする項目は以下の4つを設定した．今回，これらの項目を抽出対象に設定した理由としてこれらの項目の並列文が瀕出しているからである．

項目名	定義	記号
診療Gr分類	診断，治療の評価の分類グループ名	b
評価項目	診断，治療の評価項目，評価尺度	p
評価値	診断，治療の数値による評価値	v
評価単位	診断，治療の評価値の単位	u

表 1: 設定した抽出項目

抄録中に出現する以下のような文面における，各項目に対応する解を示す．

> 不鮮明な境界は良性腫瘤6病変と悪性腫瘍13病変で認め，周囲浸潤は良性3病変と悪性7病変でみられた。

記述例

b=不鮮明な境界,p=良性腫瘍,v=6,u=病変,p=悪性腫瘍,v=13,u=病変 b=周囲浸潤,p=良性,v=3,u=病変,p=悪性,v=7,u=病変

3 出現パターン

今回，設定した項目は各項目どうしが組となり文中に出現し，その組が1文中に複数組出現するものが多い．よって，出現パターンを調べることで，部分テンプレートを作成する鍵となる．本稿では並列表現を分析するにあたり，2000年度分573稿の医療論文抄録を用いた．

●項目どうしの出現パターン

先ほどの例における項目の出現パターンを項目の記号を用いて表すと

(1)*b-p-v-u-p-v-u-b-p-v-u-p-v-u*

となる．そこで $(x\text{-}y)?$ を x-y の組み合せが連続して出現するものと定義すると，(1) のパターンは

(1-1) *(b-(p-v-u)?)?*

と表現することができる．

出現タイプ	出現パターン	頻度
1	*b-(p-v-u)?*	218文
2	*(p-v-u)?*	126文
3	*p-v-u*	52文
4	*(b-(p-v-u)?)?*	34文
5	*p-(v-u)?*	24文
6	*(b-(p-(v-u)?)?)?*	23文
7	*(b)?-p-(v-u)?*	20文
8	*(b-(v-u-p)?)?*	19文
9	*(v-u-p)?*	18文
10	*(p)?-v-u*	12文
11	*b-(p)?-(v-u)?*	12文
12	*(p)?-(v-u)?*	10文
13	*b-(p)?-v-u*	5文
14	*p-v-u-v-u-p*	4文
15	*b-(v-u-p)?*	4文
16	*((p)?-b-(v-u)?)?*	4文
17	*(p)?-(v)?-u*	4文

表 2: 1文中の出現パターンの分類

表2に照らし合わせると先ほどの例 (1-1) は出現タイプ4に相当する．また，表2より調べた文において，並列構造となっている文は533文で90%以上の割合で並列構造を成していることがわかる．

●項目間の対応

1つの評価値 (v) に対して1つの評価単位 (u) が出現している (*v-u*)

例外 (4 文/585 文)：パターン 17 に相当する文

右室自由壁辺縁のｓｃａｌｌｏｐｉｎｇ，中隔左室側の脂肪沈着，左室自由壁の脂肪沈着はそれぞれ６，３，３例に認められた。

1 組の評価値，評価単位に対して 1 つの評価項目 (p) が出現している (p-v-u または p-p-v-u-v-u)
例外 (41 文/585 文)：パターン 5，パターン 10，パターン 13 に相当する文

左冠動脈入口部病変，左冠動脈入口部病変＋末梢病変，ともに２例確認された。

n(n ≧ 1) 組の評価項目，評価値，評価単位に対し m(n ≧ m) の診療 Gr 分類 (b) が出現している．(b-p-v-u または b-p-v-u-p-v-u)
※診療 Gr 分類は 1 文中に必ず出現するとは限らない．

と，および，や，及び，または，ならびに，:，;，また，かつ，もしくは，ないしは

4 抽出手法

先述したように医療論文抄録は限られた字数で記述しなければならないため，1 文に多くの情報が詰め込まれている．今回，抽出の対象とした項目に関しても 1 文中に多く出現している．しかし，分析結果から評価項目－評価値－評価単位といった具合に項目の並びがある程度一定しており，その組み合せが 1 文中に連続して出現しているものが非常に多い．従って，抽出方法も評価項目－評価値－評価単位の部分テンプレートを作成し，入力した 1 文に対し，それを繰りかえしマッチングさせて抽出を行った．また，部分テンプレートを作成するにあたり，評価値，評価単位に以下の情報を組み込むことで，評価値，評価単位を軸にマッチングを行うことができ，ミスマッチングを減少させ，各項目の抽出精度を向上させることができる．

評価値＝[0-9，±，−，＋，／，全]
評価単位＝[症例，病変，結節，部位，個，名，人，例，％，ｍ／ｓ，ｍｍ，ｍｏｌ …]

評価値，評価単位のパターン

部分テンプレートを作成して抽出できない例として以下のような場合がある

Ｔ病期別ではＴ１，Ｔ２，Ｔ３，Ｔ４はそれぞれ１６％，６４％，５８％，６６％で，認められた。
p=Ｔ１，Ｔ２，Ｔ３，Ｔ４
v,u=１６％，６４％，５８％，６６％

このような場合，先ほどのように部分テンプレートで評価値，評価単位を軸にマッチングさせることができない．すなわち，評価項目の並列箇所の始点と終点が決定できない．そこで構文解析器である KNP[2] の解析結果を用いて，並列箇所を特定する手法 [3] を用いた．

文章の評価項目と評価値を記述する間にあるキーワード (それぞれ，各々といった特徴語) を境にして前半部分と後半部分に分け (図 1 参照)，並列を構成している個数が同じ数のとき，解として採用する．このとき，評価値と評価単位は先ほどと同様にパターンにあてはまるものでなければならないものとした．

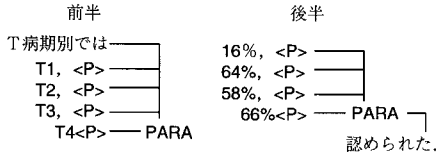

図 1: 構文解析結果

5 実験結果

実験では，1999 年度医療論文抄録 236 抄録を使用した．表 3 にその結果を示す．

抽出項目	正解	誤答	適合率	再現率
診療 Gr 分類	96	58	0.93	0.65
評価項目	553	360	0.80	0.71
評価値	641	318	0.87	0.73
評価単位	495	260	0.88	0.72

表 3: 実験結果

6 考察とまとめ

本稿では，1 文単位のテンプレートを作成するのではなく，部分テンプレートを作成し，また部分テンプレートではマッチングしないような文に対しては構文解析を用いて並列箇所を特定し，文のパターンに分けて抽出方法を変化させながら抽出を行った．1 文テンプレートを用いての抽出のときに比べ，よい結果を得ることができた．誤答のうち，そのほとんどが抽出した解の間違いによるものではなく，未抽出によるものである．よって，さらに多くの論文抄録の分析とそれに基づいた部分テンプレートの作成により，再現率の向上を目指すことができると考えられる (1 文テンプレートを用いた抽出実験では間違いの解を多く取りだしてきていた)．また，正解を得た文の項目間の対応関係はほぼとれていた．今後は，今回抽出できなかった文の分析と，多くの新しい分析データを用いて文の分析し，もっと多くの部分テンプレートを作成することで，精度向上を目指す．

謝辞

この研究は科学技術振興調整費「高度医療ネットワークに関する研究」の支援を受けて行われました。ここに深く感謝の意を表します．

参考文献

[1] 井上 大悟，永井 秀利，中村 貞吾，野村 浩郷，大貝 晴俊：医療論文抄録からのファクト情報抽出を目的とした言語分析，情報処理学会研究報告 01-NL-141,pp.103-110 2001

[2] 黒橋 禎夫，日本語構文解析システム KNP 使用説明書 version2.0 b6，京都大学大学院工学研究科

[3] 赤松 順子，永井 秀利，中村貞吾，野村浩郷：複数製品の紹介記事からの製品情報抽出，情報処理学会研究報告 00-NL-140,pp.61-68 2000