自然言語処理 141－17
（2001．1．26）

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

井上 大悟　永井 秀利　中村 貞吾　野村 浩郷　‡大貝 晴俊

九州工業大学大学院 情報工学研究科
‡科学技術振興事業団
E-mail: {inoue,nagai,teigo,nomura}
@dumbo.ai.kyutech.ac.jp
‡ E-mail: ogai@tokyo.jst.go.jp

我々は従来テンプレートを用いて新聞記事からの情報抽出の研究を行ってきた．そこで今回対象を医療分野に広げ，医療論文抄録からの情報抽出を行う．
医療論文抄録には多くのファクト情報が含まれており，我々は45種類の項目を設定した．本稿ではその中から最も重要なファクト情報である病名，診断機器，診療症例，診断結果の項目に対して記述パターンや文中に共起する文字列について分析した．
それを基に抽出すべき情報とその周辺の文字列との関係を記した"テンプレート"を用いて字面処理による抽出実験を行い，その結果を報告する．

Analysing Description Patterns for Information Extraction from Abstracts of Medical Articles

Daigo Inoue　Hidetoshi Nagai
Teigo Nakamura　Hirosato Nomura
‡ Harutosi Oogai

Graduate School of Information Engineering, Kyushu Institute of Technology
‡ Japan Science and Technology Corporation(JST)
E-mail: {takao,nagai,teigo,nomura}
@dumbo.ai.kyutech.ac.jp
‡ E-mail: ogai@tokyo.jst.go.jp

We have been researching product information extraction from newspaper articles using template matching. We would extend the target domain to investigate the method of information extraction, and chose medical articles as the new domain. Medical articles contain various information about medical treatments and we set up 45 kinds of item to be extracted from abstracts of medical articles. In this paper, we select some important items out of those and analyze these description patterns and cooccurrence expressions in abstracts of medical articles. We show the result of extraction experiment using templates which describe the relationship between information to be extracted and its surrounding strings.

1 はじめに

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，倫理，制度，技術などを含め医療のこれからのあり方自体を根本的に改革される時期に差し掛かっている．医療の進歩とコンピュータ，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ため，医者や技師がそれらの大量の文書の中からファクト情報を手作業だけで取り出してくることは，とても困難になりつつある．このような状況下において，カルテや医療文献からファクト情報を計算機で自動的に処理し，取り出してくることができれば，情報を効率的に管理することができ，また更新された情報もすばやく伝えることができる．[7] このように，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．[5][6]

筆者らは，大量の計算機可読文書から目的とする情報を抽出する研究を行っており，これまでに新聞記事を対象として文章の記述形式を分析し目的とする情報に対し，どのような抽出方法がよいかを模索してきた．文書が定型性を持ち，かつ抽出対象が明確である場合には，構文解析や意味解析を行わずとも字面処理による簡易な処理で情報を高速に抽出することができる．そこで医療論文抄録は制限された字数で表現されなければならないため，無駄を省いた表現が多く，文体も一定している．このような場合，情報抽出する手法として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"を使用する方法がある．[1][3][3][4]

そこで，医療論文抄録から情報抽出を行うことができるように，抄録中の目的とする情報に対しタグ付けを行い，それを基に記述パターンを分析した．そして，分析した結果を基に各項目の字面や字種の特徴を生かして医療論文抄録からのファクト情報抽出の実験を行った．

2 医療論文抄録

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．そのうち692稿に対してファクト情報部分にタグ付けを行い，論文抄録の分析を行った．医療論文抄録はいずれも

1. 論文番号　タイトル
2. 所属　発表者氏名
3. 本文

の形式で記述されており，その情報を表1に記す．

総論文数	タイトル平均文字数
930	32.5

1抄録の本文			本文1文 平均文字数
総文数	平均文数	平均文字数	
9479	10.2	604.0	59.3

表 1: 医療論文抄録に関するデータ

今回分析した論文抄録を大別すると以下のようになる．

・一個人の症例の報告
・病気の発生状況を統計データとして報告
・病気の治療法とその効果の報告
・各種の医療機器により病気の診断方法の報告

2.1 医療論文抄録の文章構成

今回用いた論文抄録の本文中にはそれぞれの文章構成を示すため，インデキシングがされている．付録の例では,【目的】【方法】【結果】【結論】がそのインデックスの部分にあたる．そこで，1999年度分236稿についてインデックスの種類とその出現回数を調べた．表2に記す．

意味的に同義の語を統一した結果より，医療論文抄録は 目的 ，方法 (対象) ，結果，結論 (考察) と大きく4つに分けて構成されていることがわかる．

2.2 抽出項目の設定

医療論文抄録は多くのファクト情報を含んでおり，専門家にヒアリングして目的とする情報を定義したところ項目数は45にも及んだ．定義した項目の一部を表3に記す．それらの中で，我々は抽出を試みる内容である「抽出項目」を論文抄録中に出現する主なものである病名，診断機器，診療対象症例，診療症例例数，診療症

表現毎の分類		同義表現を統一	
インデックス	出現数	インデックス	出現数
目的	226	目的	229
結果	223	結果	228
結論	160	方法	226
方法	113	結論	221
対象と方法	57	対象	144
結語	44	考察	26
対象	33		
対象および方法	20		
対象・方法	12		
考察	12		
対象及び方法	6		
対象ならびに方法	6		
まとめ	4		
…	…		

表 2: インデックスの種類と出現回数

例例数単位と診断結果について述べている診断分類，診断性能，性能単位に設定し，各抽出項目に関して記述表現，項目自身がもつ特徴を分析した．

記号	抽出項目
RB	論文番号
RM	論文題目
HM	発表者氏名
HS	発表者所属
BM	病名
SK	診断機器
TK	診療対象症例
TR	診療症例例数
TU	診療症例例数単位
SNB	診断分類
SNV	診断性能
SU	性能単位

表 3: ファクト情報種別

3 抽出項目の記述形態の分析

3.1 病名

病名は「論文が取り扱っている主病名」と定義している．先に記述したインデックスにおいて，題目，目的，方法・対象，結果，結論と論文抄録を5つに大別した場合，病名の出現箇所と出現数は表4のようになった．

表4はタグ付けした抄録692稿を調べたものである．従って，1抄録当たり平均1.1となり，

出現箇所	出現数
題目	212
目的	464
方法・対象	77
結果	0
結論	7
合計	762

表 4: 病名の出現箇所とその数

ほぼ1抄録に1回しか出現していないことがわかる．

病名の字種的特徴

病名の多くは身体組織や病名修飾語や病名接尾辞，病名と病名が組合わさった複合語である．病名の接尾辞は特徴的であり調べてみたところ以下のものが大多数を占めた．

接尾辞: ～症，～炎，～腫，～群，～病，～毒，～血，～癌，～がん，～狭窄，～病変，～疾患，～瘤，～瘍，～塞，～障害，～傷

接尾辞以外の部分での病名の字種パターンは5通りある．

接尾辞以外:

●アルファベット列 (24個)
例) Ｐａｒｋｉｎｓｏｎ病，Ｂａｓｅｄｏｗ病
●片仮名列 (13個)
例) シェーグレン症候群，アルツハイマー病
●漢字列 (285個)
例) 肝細胞癌，外傷性胸部大動脈瘤
●上記3つのミックス型 (17個)
例) 眼窩原発非ホジキンリンパ腫，乳房外Ｐａｇｅｔ病
●その他 (3個)
例) びまん性脳損傷，びまん性肺疾患

病名と共起する特徴ある文字列はみられなかったが，病名自身がもつ記述パターンを作成することにより抽出が可能と考えられる．

3.2 診断機器

診断機器は「診断または治療で使用した機器名称」と定義している．先ほどと同様に論文抄録を5つに大別した場合，診断機器の出現箇所

は「方法・対象」の部分に約97％の割合で出現している．(表5を参照)

出現箇所	出現数
題目	4
目的	9
方法・対象	405
結果	0
結論	0
合計	418

表 5: 診断機器の出現箇所とその数

診断機器の字種的特徴

診断機器も病名と同様に構成する字種がほぼ限られている．分析用データから得られた診断機器の字種構成を調べたところ片仮名，漢字，数字，記号，アルファベットの組合せからなるものが417個で1個だけ平仮名を含むものであった．
また，診断機器は ～(社) 製～ という文字列を含むものが69個あり，この文字列を含み，かつ先に記述した字種構成の文字列は診断機器として抽出することができる．

1文中に診断機器と共起する文字列

診断機器を含む1文は非常に特徴的で，特に文の開始や項目の直後に頻出する文字列がある．以下に代表的なものとその数を記す．
例)～装置は { 診断機器 }～(76)
～{ 診断機器 } を用い～ (76)
使用した～{ 診断機器 }～(27)
～{ 診断機器 } を使用し～ (22)
～機種は { 診断機器 }～ (21)
～{ 診断機器 } を～撮像した (18)
これらの記述形式の定型性を利用して，テンプレートによる抽出が可能と考えられる．

3.3 診療対象 (症例・例数・単位)

診療対象症例は「診断，治療の対象とした患者，症例」と定義しており，診療を対象とした人 (症例) の数として診療症例例数，その単位として診療症例例数を定義している．また，これらの項目は方法・対象の部分にしか出現しない．出現数は表6となっている．各数値が微妙に異なるのは単位の省略や並列構造の記述形式が原因である．

症例	例数	単位
826	833	831

表 6: 診療症例の出現数

診療対象の字種的特徴

診療対象症例を構成する字種に目立った特徴はみられなかった．しかし，診療対象例数は数値で表記され，その直後にくる単位 (診療症例例数単位) には以下のものがあった．
～症例，～人，～名，～頭，～匹，～例，～結節，～病変，～体，～枚

1文中に項目と共起する文字列

診断機器と同様に項目を含む1文の開始や文中，文末に頻出する文字列がある．
例) 対象は { 症例 }{ 例数 }{ 単位 }～(163)
{ 症例 }{ 例数 }{ 単位 } を対象とした．(153)
{ 症例 }{ 例数 }{ 単位 } を対象とし，～(54)

抽出項目どうしの対応関係

対象は {TK}HCC{/TK}{TR} １０ {/TR} {TU} 症例 {/TU} で，腫瘍径３ｃｍ以下である。

上記のように，診療対象症例，例数，単位は同じ1文中に出現し，そのほとんどが1抄録中に1回しか出現しないが，以下のような場合もある．

対象は {TK} Ａｌｚｈｅｉｍｅｒ型痴呆症例 {/TK}{TR} ８ {TR}{TU} 例 {/TU} と {TK} 非Ａｌｚｈｅｉｍｅｒ症例 {/TK} {TR} ８ {/TR}{TU} 例 {/TU} である。

このように複数回出現し並列構造になっている場合もあるが，診療対象症例-例数-単位と抽出項目が出現するパターンは一定しており，パターンを記述することで対処できる．これらの特徴を考慮すると診療対象項目もテンプレートによる抽出が可能と考えられる．

3.4 診断結果 (分類・性能・単位)

ここでは，診断した結果の診断分類，診断性能，性能単位について述べる．これらは結果のインデックス中に出現し，それぞれ診断分類は

「診断，治療の評価の分類」，診断性能は「診断，治療に対しての性能評価」，性能単位は「診断性能が数値で表された場合の単位」と定義している．

診断結果の字種的特徴

診断分類を構成する字種に目立った特徴はみられなかった．診断性能，性能単位と隣接して出現する場合，診断性能は数値もしくは"全"であり，単位は
～症例，～例，～%，～結節，～cm，
～病変，～mm，～cm／s，～ml／min，
～分，～秒
など計33種類あった．

1文中に診断結果と共起する文字列

診断結果では項目を含む1文に特徴ある文字列が頻出する．
例)～であった．(274)
～描出された,～描出した(136)
～示した，～示された(106)
～認められた．(72)
～を認めた(38)

抽出項目どうしの出現パターン

抄録の記述を見ると，診断分類の診断性能に対する出現パターンが異なる場合がある．多くは下記A，Bのような，診断分類と診断性能がn対nのものが多いが，Cのような場合もある．

A：1(診断分類) 対 1(診断性能)

{SNB}経時的差分画像{/SNB}は{SNV}93.4{/SNV}{SU}%{/SU}でアーチファクトの少ない診断に適した画像が得られた。

B：n(診断分類) 対 n(診断性能) n ≧ 2

初再発から経過中に{SNB}腫瘍内の動脈血流が増加{/SNB}したと思われるものが{SNV}6{/SNV}{SU}結節{/SU}，{SNB}門脈血流が低下{/SNB}したと思われるものが{SNV}3{/SNV}{SU}結節{/SU}に認められた。

C：m(診断分類) 対 n(診断性能)(m ≠ n)

{SNB}ＦＳＥ法Ｔ２強調画像{/SNB}と{SNB}ＦＥ法Ｔ２＊強調画像{/SNB}はいずれも{SNV}４０{/SNV}{SU}結節{/SU}の描出が可能であった

各パターンの出現頻度は表7を参照．

Aパターン	Bパターン	Cパターン
170文	141文	40文

表7: 各パターンの出現数(1999年度236稿)

A，Bパターンのように分類-性能-単位と出現する順が一定しており診療症例と同様にテンプレートによる抽出が可能と考えられる．ところが，Cパターンのように並列構造をともなって記述されているものがあり，省略が発生しているものも多くみられる．

4 テンプレートを用いた情報抽出

4.1 テンプレート

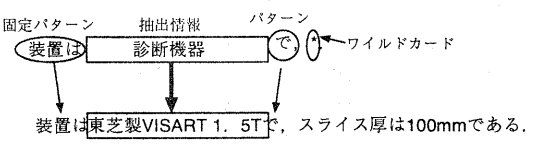
今回分析した結果をもとにテンプレートによる抽出を行う．テンプレートとは抽出する項目とその周辺の文字列を記述したもので，テンプレート作成に用いる用語を以下に示す．

抽出情報: 抽出項目に対応する情報を表すラベル

パターン: パターンマッチングの対象となる文字列長1文字以上の文字列

固定パターン: 抽出対象に頻出する特徴的な文字列("対象は"，"装置は")

ワイルドカード: パターンマッチング上，文字列長0以上の任意の文字列とマッチしうるシンボル



L を抽出項目，P をパターン，W をワイルドカードとしたとき，テンプレート T を

$$T = C_0L_1C_1L_2\cdots C_{n-1}L_nC_n$$
$$(C_i = P_0W_1P_1W_2\cdots P_{m-1}W_mP_m)$$

と定める．なお，文頭の C_0 および文末の C_n は空文字列であってもよい．テンプレートは1文単位で作成し，C と L は必ず交互に現れるものとする．

テンプレートの優先順位付け

ここではテンプレートを決められた順番でマッチングさせていき，マッチングが成功した時点で処理を終了する方法を取ることにする．テンプレートの並びを決定する方法は以下の通り．

1 抽出システムを用いて各テンプレートでパターンマッチングを行い，テンプレート作成用データのすべての文から情報抽出を行う

2 テンプレートとマッチした文の数を求めて，数が少ないテンプレートから順番にテンプレートを並べかえる．

4.2 抽出実験システム

テンプレートを用いた情報抽出実験システムについて述べる．システムの実装にあたってスクリプト言語を使用した．本システムの概要図を図1に示す．

各モジュールの処理

入力文加工モジュール

入力された論文抄録中で抽出する項目が頻出するインデックス部分を取り出し，文単位に分割する．

パターンマッチングモジュール

入力文加工モジュールで分割された1文を入力として受けとり，テンプレートとのマッチングを行い，マッチングによって抽出に成功したすべての情報を出力とする．

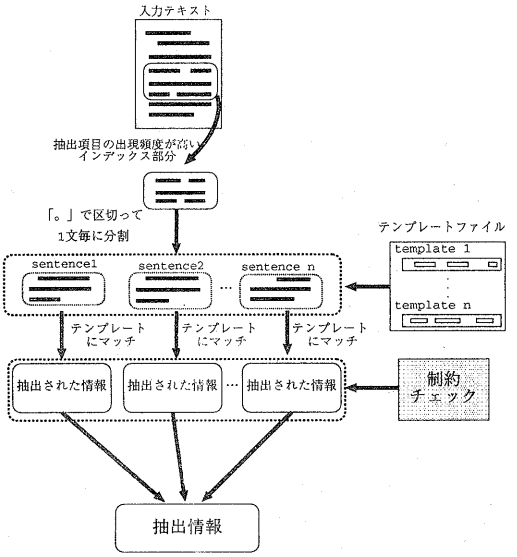


図 1: テンプレートを用いた情報抽出システム

制約判定モジュール

制約判定モジュールは，パターンマッチングモジュールで得た文字列を渡され，制約判定の合否をパターンマッチングモジュールに返す．制約判定の結果が1つでも失敗すれば制約判定モジュールは不合格であり，パターンマッチングモジュールに戻って先ほどマッチしたテンプレートの次のものとマッチングを行う．

- 括弧の対応がついていなければ失敗
- 禁則開始文字で始まるならば失敗
- 読点から始まるならば，読点以降の文字列を返す．

5 実験と評価

実験では，医療論文抄録1999年度，2000年度分802稿を使用した．そのうちタグ付けした抄録692稿をテンプレート作成用データとし，タグ付けされてない110稿を評価用データとした．病名の抽出には抽出情報自身がもつ記述 (字種) の表記パターンを作成して抽出実験を行った．病名以外の項目には，4章で説明し

たテンプレートを作成して抽出実験を行った．評価方法は再現率，適合率を用いた．

5.1 実験結果

病名

病名は論文で扱われる主病名と定義してあるため，題目に出現するものを優先的に抽出し，題目に出現しない場合はインデックスの目的部分から抽出ように設定した．

抽出項目	適合率	再現率
病名	0.90(88/97)	0.98(88/90)

表 8: 病名抽出結果

診断機器

診断機器は方法・対象のインデックス部分の文に対し抽出処理を行った．テンプレートの数を増やすことで精度の向上が可能と考えられる．

抽出項目	適合率	再現率
診断機器	0.87(55/63)	0.76(55/72)

表 9: 診断機器抽出結果

診療症例

診断機器と同様に方法・対象のインデックス部分の文に対して処理を行った．結果は表 10 となっている．診療症例例数，診療症例例数単位は表記が決まっており，テンプレートに表記パターンを加えることでよい結果を得た．診療対象症例に関してはテンプレートの並びに大きく左右され，1 番のぞましいテンプレートにマッチしない場合があった．

抽出項目	適合率	再現率
診療症例	0.87(95/109)	0.77(95/123)
診療例数	0.93(108/116)	0.88(108/123)
診療単位	0.95(109/115)	0.89(109/122)

表 10: 診療症例抽出結果

診断結果

診断結果部分は結果のインデックス部分の文に対して処理を行った．結果は表 11 となっている．

抽出項目	適合率	再現率
診断分類	0.63(195/309)	0.50(195/384)
診療性能	0.51(347/677)	0.61(347/572)
診療結果	0.86(238/276)	0.72(238/329)

表 11: 診断結果抽出結果

結果が悪い原因として並列構造の文が多さが挙げられる．記述形式も筆者により異なるためテンプレートを作成しての抽出ではよい結果は得られなかった．医療論文抄録では字数が制限され，無駄を省いた表現が多く，項目を読点や「と」,「や」といった並列キーで情報を列挙して記述されており，項目間の対応関係がうまくとれなかった．

6 まとめ

医療論文抄録にタグ付けし，それを基に記述パターン，抽出項目自身がもつ字種パターンを分析し，試験的に抽出実験を行った．

抽出項目周辺の文字列であるテンプレートを入力した文とのマッチングだけでなく，抽出項目自身がもつ字種パターンの利用により，よい結果を得ることができた．

しかし，診断結果部分については項目自身が字種パターンをもたず，並列構造の文が多く出現するためテンプレートによる抽出ではよい結果を得られなかった．並列構造のテンプレートの数を増やして実験を行い，テンプレートを用いた抽出が有効であるか確認する必要がある．

謝辞

科学技術振興事業団には，今回用いた実験データの提供やテキストへのタグ付け作業など色々とお世話になりました．この場を借りて深く感謝します．

参考文献

[1] 高尾 宜之，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 -製品記述パターンの分析-，情報処理学会研究報告 99-NL-129, pp. 117-124, 1998

[2] 井出 裕二, 藤吉 誠, 永井 秀利, 中村貞吾, 野村浩郷: テンプレートを用いた新聞記事からの製品情報抽出システム, 情報処理学会研究報告 96-NL-115, pp. 83 – 90, 1996

[3] 井出 裕二, 藤吉 誠, 永井 秀利, 中村貞吾, 野村浩郷: 構造化テンプレートを用いた新聞記事からの製品情報抽出, 情報処理学会研究報告 97-NL-118, pp. 7 – 14, 1997

[4] 井出 裕二, 永井 秀利, 中村貞吾, 野村浩郷: 単一項目テンプレートによる新聞記事からの製品情報抽出, 情報処理学会研究報告 97-NL-122, pp. 63 – 70, 1997

[5] 江里口 善生, 木谷 強: 冨田一般化LRパーザを用いた情報抽出, 情報処理学会論分誌 Vol.38 No.1,pp44-54,1997

[6] 江里口 善生, 木谷 強: 冨田一般化 LR パーザを用いた情報抽出, 情報処理学会研究報告 94-NL-102,pp.9-16,1994

[7] 岡野 弘行, 大杉 英男, 大貝 晴俊: 情報の収集技術の研究, 平成 10 年度科学技術振興調整費 調査研究成果報告書

7 付録

1 冠動脈バイパスグラフトの高速MR血流量計測：吻合部狭窄病変の非侵襲的診断は可能か？

三重大・放 河田七香, 佐久間肇, 野村新之, 加藤憲幸, 竹田 寛
三重大・胸外 Ｂａｙａｒｄｏ Ｐ Ｃｒｕｚ
松阪中央病院・放 平野忠則

【目的】高速ＭＲＩを用いて内胸動脈（ＩＭＡ）－冠動脈バイパスグラフトの血流量を計測し，グラフト狭窄の非侵襲的診断の可能性を検討する。【方法】対象はＸ線アンギオ上グラフト開存の確認されたＩＭＡグラフト患者２３例である（Ａ群１８例：Ｘ線アンギオ上正常または吻合部狭窄＜７５％，Ｂ群５例：吻合部狭窄＞７５％）。高速ｖｅｌｏｃｉｔｙ ｅｎｃｏｄｅｄ ｃｉｎｅ ＭＲＩを用いてジピリダモール負荷前後におけるＩＭＡグラフトの血流計測を行った。画像解析にはＸｐｈａｓｅを用い，血流量と拡張期／収縮期ピーク血流比（Ｄ／Ｓ ｒａｔｉｏ）を求めた。【結果】Ｂ群における負荷前グラフト血流量（１４．１±１１．０ｍｌ／ｍｉｎ）はＡ群（７６．４±３８．３ｍｌ／ｍｉｎ）と比較して有意に低下していた（ｐ＜０．０１）。安静時ＭＲ血流量計測によるグラフト狭窄診断のｓｅｎｓｉｔｉｖｉｔｙは１００％，ｓｐｅｃｉｆｉｃｉｔｙは８９％であった（閾値３５ｍｌ／ｍｉｎ）。また，Ｂ群のＤ／Ｐ ｒａｔｉｏ（０．９８±０．６７）はＡ群（２．０１±１．３７）より有意に低く（ｐ＜０．０１），Ｄ／Ｐ ｒａｔｉｏに基づくグラフト狭窄診断のｓｅｎｓｉｔｉｖｉｔｙ，ｓｐｅｃｉｆｉｃｉｔｙはそれぞれ１００％，８９％であった。一方，グラフト血流予備能にも両群間に有意差が認められたが（Ａ群２．０１±１．３７，Ｂ群０．９９±０．６６，ｐ＜０．０５），２群の分布にはかなりの重なりがみられた。【結論】高速ＭＲを用いたＩＭＡグラフトの血流計測は有意狭窄病変の非侵襲的診断に有用であり，Ｘ線アンギオに代わり得ると思われる。グラフト狭窄の検出を目的としたＭＲ血流計測に薬物負荷は必ずしも必要ではない。