社団法人　情報処理学会　研究報告
IPSJ SIG Technical Reports

2005－ＮＬ－168（13）
2005／7／23

医療用語のコード化手法
－部分文字列のテキスト近似被覆問題－

藤尾 正和† 関 峰伸† 永崎 健† 丹羽 芳樹† 久光 徹†

†（株）日立製作所 中央研究所 〒185-8601 東京都国分寺市東恋ヶ窪一丁目280番地
E-mail: †{m-fujio,mseki,naga-t,yniwa,hisamitu}@crl.hitachi.co.jp

概要 本研究では，大量の診療報酬請求書（レセプト）の審査業務を効率化することを目的として，レセプト記載文字列を統制用語にコード化する手法を提案した．文字認識結果をコード化するには，元レセプトに存在する表記のゆらぎに加え，誤字脱字の存在を前提としてコード化を行う必要がある．本研究ではこの問題を，入力テキストを統制用語の部分文字列で適切に被覆する問題と考え，動的計画法により最小コスト被覆を求める実験を行った．500枚（約18,000行）のサンプルで評価した結果，正解テキストを用いた場合で84%，文字認識後の誤読・不読データを用いた場合で65%のコード化率を達成し，手法の有効性を確認した．

キーワード　レセプト，OCR，表記ゆれ，コード化，医療用語，被覆

Medical Treatment Encoding Method for Clinical Data Analysis
From Medical Care Bills
－Text mapping by the substrings extracted from controlled terms－

Masakazu FUJIO† Minenobu SEKI† Takeshi NAGASAKI† Minenobu SEKI†
Yoshiki NIWA† and Toru HISAMITSU†

† Hitachi, Ltd., Central Research Laboratory 1-280, Higashi-koigakubo, Kokubunji-shi, Tokyo, 185-8601, Japan
E-mail: †{m-fujio,mseki,naga-t,yniwa,hisamitu}@crl.hitachi.co.jp

Abstract This paper propose a method to transform medical treatment text into set of controlled terms, by way of substrings mapping. Those substrings are extracted from the controlled terms. The proposed method re-arranges the subset of substrings over the input text in minimum cost manner, which will be defined later. By applying the method to 500 medical care bill images, we achieved 84% precision for human transcription of the same samples, and 65% precision for OCR output from the images.

Keyword recept, ocr, approximate string matching, dynamic programming

1. 背景

整備済みの辞書を用いる自然言語処理の多くのアプリケーションでは，様々な表記ゆれの存在が実用面での大きな障害となっている．例えば，バイオメディカル分野の文献では，タンパク質名の省略やニックネーム，また別々の研究者により独立に発見された同一タンパク質の存在が，相互作用情報抽出などのタスクを著しく困難なものとしている．また，医療機関で作成されるカルテや読影レポート，毎月の診療報酬請求書等の書類も，医師・診療科により用語や表現が多種多様であり，機械処理を困難なものとしている．そのため，バイオメディカル分野では，UMLS(Unified Medical Language System)，SNOMED，SNOMED-CT (Systematized Nomenclature of Medicine-Clinical Terms)，Gene Ontology，Mesh Term，ICD10（国際疾病分類）など，辞書・シソーラスの整備が精力的に行われている．

医療機関で扱われる文書の中でも比較的書式が固定し，処理の機械化が進んでいるものとして，診療報酬請求書（レセプト）が知られている．

レセプトとは，医療機関において毎月の診療行為・薬剤・医療材料等の請求内容を患者単位にまとめた診療報酬請求書のことを言う．医療機関は，レセプトを元に，健康保険組合などに，医療費の請求を行う．レセプトの流通量は，年間12億～14億枚と言われており，厚生労働省・社会保険庁では事務処理の効率化・迅速化を目的として，レセプト電算処理システムの普及を目指している．その一環として，診療報酬請求に用いるべき統制用語（マスター）を定義し，固有のコードを付与して管理している．しかし，医科部門での普及率はレセプト件数ベースでは依然 3%以下に留まっており（2003年12月現在：社会保険診療報酬支払

基金調べ[1]，審査業務の大部分が医療機関から紙で提出された請求書もしくはスキャンした画像を目視することで行われている．

筆者らは，審査支払機関や保険者において行われる審査業務の効率化するため，レセプト画像データのOCR出力を統制用語に変換（コード化）する方法を検討してきた．レセプトは通常文章ではなく，行単位で一つ以上の請求項目がリストされたものであり，またコード化する必要のない記述もある．従って，画像データをコード化するためには，レセプトに元から存在する表記ゆれのみならず，OCR誤読や不読による表記ゆれのあるテキストから，請求内容を過不足なく抽出する技術が求められる．

本研究では，入力テキストを統制用語の部分文字列により適切に被覆する問題と捕らえ，動的計画法による最小コスト探索を応用してこの問題を解く実験を行った．以降まず2章で，レセプト記述の特徴について述べたあと，3章でコード化手法について説明する．4章で，500件のレセプトを用いた評価実験と，事例分析，考察について述べ，5章で本タスクに関連するもしくは有効と思われる研究について紹介する．最後に，6章で，まとめと今後の課題について述べる．

2. レセプト摘要欄の特徴

2.1. コード化における課題

文字認識結果を用いる本タスクでは，統制用語マスター[2]で定義されている診療報酬請求用語と，実際のレセプトに記載されている処置・処方名との間の表記ゆらぎのほか，文字認識結果の誤読・不読による表記ゆらぎが存在する．

2.1.1. 表記のゆらぎ

レセプト摘要欄の請求項目は，主に診療行為，調剤行為，薬剤名，医療材料から構成されており，請求内容を表さないコメントが含まれることもある．統制用語マスターでは，これらの項目は，括弧やスペース，点を含んだ複合語を構成していることが多い．例えば，診療行為であれば，「初診時(診療所)紹介患者加算」，「精密眼底(片)」，調剤行為であれば，「処方せん料(6種類以下)（後発医薬品を含まない)」，薬剤名では，「メイアクト錠１００　３錠」などが，統制用語マスターとして存在する．実際のレセプトでは，括弧が省略されたり，括弧の中と外が入れ替わったり，“錠”という語が省略されたり，容量単位が異なったり，もしくは医学的には同義であるまったくべつの用語が使われたりする．

例えば，「Ｎａ，Ｃｌ，Ｋ，血糖」という表記は，「Ｎａ及びＣｌ」，「Ｋ」，「グルコース」という3つの統制用語マスターにコード化されなくてはならない．また，「精密眼底（片)」は「精密眼底検査（右側)」などと書かれることもある．また，「心電図検査（１２誘導)」とは，統制用語マスターで言う所の「ECG１２」であり，「療養担当手当（外来)」は，「療養担当手当（病院外)」のことである．省略表現では，「薬剤情報提供量」が「薬情」，「尿沈渣顕微鏡検査」が「尿沈」などの表現が見られる．

表 1に，摘要欄において見られる代表的な表記ゆれ（文字認識誤りについては除く）についてまとめた．

表 1：摘要欄にみられる表記ゆれの実例

表記ゆらぎの種類	レセプト中の表記（文字認識後）
カナ/漢字	・「処方せん」、「処方箋」 ・「タンパク」、「蛋白」、「たんぱく」
ローマ数字、機種依存文字	・「紹介患者加算（4)」、「紹介患者加算４」、「紹介患者加算（ⅠV)」、「紹介患者加算（Ⅳ)」
数字	・「１８，０００」、「１８０００」 ・「１．２０」、「１・２０」
略語	・「総タンパク」、「ＴＰ」　・「血液採取料（静脈)」、「Ｂ－Ｖ」　・「ヘモグロビン」、「Ｈｂ」
省略	・「Ａｍｙ（尿)」、「アラーゼ」　・「尿酸測定」、「尿酸」 ・「再診料（病院)」、「再診（病院)」
語順違い	・「大塚生食注２０ｍＬ」、「生食注（大塚）２０ｍＬ」
異体字	・「頚」、「頸」　・「灯」、「燈」　・「膣」、「腟」
誤字・脱字	・「尿・糞便等顕微鏡検査」、「尿・糞便顕微鏡検査」
同義語	・「ＧＴＰ」、「Ｇ・Ｔ・Ｐ」、「グルタミック・ピルビック・トランスアミナーゼ」 ・「活性化部分ＰＰＴ」、「活性化部分トロンボプラスチン測定」 ・「スリットＭ」、「細隙燈顕微鏡検査」 ・「特定疾患療養指導料（病院)」、「特定疾患療養指導料（１００床以上２００床未満)」 ・「処方料（６種類以下)」、「処方料（その他)」 ・「レチクロ」、「網赤血球数」
言い換え	・「Ｎａ及びＣｌ」、「Ｎａ，Ｃｌ」

2.1.2. 多義性

同じ表記であるが，別の意味を表す例が見られる．例えば，「カプセル」のことを「Ｃａｐ」と書く場合があり，表記ゆれへの対処から，大文字と小文字の区別をしない場合，コード番号「160027210」の「ＣＡＰ」と区別が付かなくなる．また，薬剤名コードの中には「～錠　２０ｍｇ」などのように，処方量が含まれることがあるが，正規化処理によって，「ｍｇ」の部分が，血液検査項目の一つである「Ｍｇ」（コード番号は160022210）と表記が同じになってしまう．

2.1.3. 一項目の区切りの曖昧性

レセプトの各行は，マスターの一項目に一対一に対応するとは限らず，一対多の場合も多い．例えば「生化学的検査判断料」という項目が出現した直後には，次の区切り文字が出現するまでは，「γ－ＧＴＰ」，「クレアチニン」，「ＧＯＴ」，「尿酸」等の実際の検査細目が連続する傾向がみられる．逆に，薬剤名や処方名が

[1] Web版月刊基金：http://www.ssk.or.jp/web/index.html

[2] 診療報酬情報提供サービス，http://202.214.127.149/

長いものの場合，本来一行に一項目書くべきものが，複数行に分かれてしまうこともある．

2.2. 問題定義

まとめると，レセプト摘要欄の特徴は，

1） 各行は，診療行為，調剤行為，薬剤名，医療材料名の羅列
2） 行中の項目数はあらかじめ不明
3） 多種多様な表記ゆれ，コード化と無関係な文字列が存在する

ということになる．

本研究では，レセプト記載項目のコード化問題を，「入力テキスト T を，マスターの部分文字列集合との近似性に基づき，適切な被覆を構成する問題」と捕らえることにする．

3. コード化手法

処理の全体概要を，図 1に示す．大きく分けて，前処理と，コード化処理本体とに分けられる．前処理は，同義語による統制用語の拡充と，部分文字列への分解処理から構成される．コード化処理は，入力行の正規化処理と，提案手法によるコード化処理本体とに分けられる．

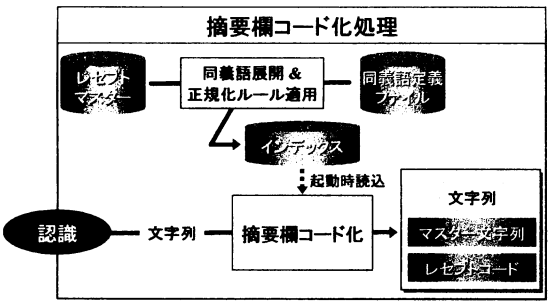

図 1：コード化処理概要

3.1. 前処理(同義定義による辞書の拡充)

統制用語辞書を，あらかじめ作成した同義語定義ファイルに従い拡張する．例えば，「入院外」⇔「外来」という同義語定義がされていると，統制用語マスター「療養担当手当（入院外)」(199000610）から，「療養担当手当金(外来)」という拡張マスターが生成される．統制用語として，平成 16 年度改定版の，診療行為マスター（約 5500 エントリー)，調剤行為マスター（約 100 エントリー)，特定保険医療材料マスター（700 エントリー)，医薬品マスター（約 17500 エントリー）を用いた．同疑語辞書の構築にあったっては，レセプト略語集[1]や，医科診療報酬点数表[2]を参考とした．頻出同義語の登録を行った．同義語の定義は，例えば「入院外」と「外来」など，マスターの部分文字列単位で定義できるようにし，同義語登録を効率的に行った．表 2に，今回用いた同義語定義の一部を示す．

表 2：同義語登録事例

種類	同義語		
同義語	直腸鏡	E-直腸	
同義語	Ｈｂ	ヘモグロビン	ヒトヘモグロビン
同義語	処方せん料（７種以下）	処方せん料（７種類以下）	処方せん料（その他）
同義語	処方料（７種以下）	処方料（７種類以下）	処方料（その他）
同義語	ＧＰＴ	Ｇ・Ｐ・Ｔ	
切断	活性化部分	活性化　部分	
切断	注射用蒸留水	注射用　蒸留水	
同義語	単純撮影（イ）	単純撮影（胸部）	
同義語	単純撮影（イ）	単純撮影（頭部）	
同義語	単純撮影（イ）	単純撮影（耳）	
同義語	単純ＣＴ撮影（ロ）	単純ＣＴ撮影（頭部）	
同義語	単純ＣＴ撮影（ロ）	単純ＣＴ撮影（躯幹）	

次に，拡張したマスター辞書を，分解規則に従い部分文字列に分解する．部分文字列への分解は，括弧の有無，数量文字列判別，ハイフン，スペースの有無，文字種の切れ目を元にした．但し，「Ｓ－Ｍ」，「Ｃ３」等，分解しないことが適当と思われるパターンについては，分解を行わないこととした．これは，後に述べる提案手法の性質から，語順を変えた，「ＭＳ」や「３Ｃ」にマッチする可能性があるためである．各部分文字列は，その種類に応じて重み付けを行う．図 2に，統制用語の分解例をいくつか示す．

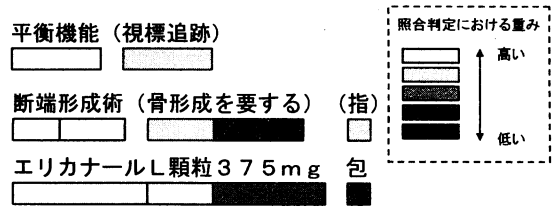

図 2：辞書の分解例

作成された部分文字列は，由来するマスターのコード番号（通常複数のコード番号を持つ）情報とともに適切なデータ構造により保持しておく．以降，拡張・分解処理によって作成された統制用語の部分文字列辞書のことを，インデックス辞書と呼ぶことにする．

3.2. 部分文字列を用いた入力テキストの被覆

コード化対象となる入力文字列に対し，まず正規化処理を行う．正規化処理では，半角はすべて全角に直し，小文字も大文字に変換している．括弧，ハイフン，句読点も統一した．正規化済みの入力文字列から，前節で作成したインデックスを要素とするラティスを構成する．図 3に，入力文字列「初診紹介患者加算（ＩＶ)」に対するラティス構造の例を示す．

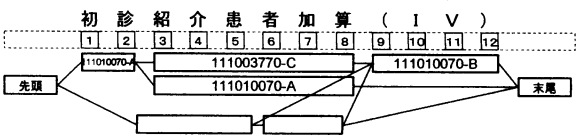

図 3：初期ラティス構造

図中の番号つきの四角は，前節ステップにより分解されたマスター部分文字列を表し，ラティスのノードとなる．例えば，文字列「初診」は，コード番号 111000111 の部分文字列であることを表し，「紹介患者加算」はコード番号 111003770 および 111010070 の両方の構成要素であり，「ＩＶ」が，コード番号 111010070 の構成要素であることを表している．各ノード間を結ぶ線がラティスのエッジである．本手法で構成するラティスにおいて，エッジで繋がれるノードは，入力テキスト上で隣接している必要はなく，Gap を許すものとする．本手法で構成するラティスの一つのパスは，一つの被覆解に対応している．

次に，ラティスの拡張処理を行う．ラティス上で隣接する部分文字列のうち，由来するコード番号集合の間の積集合を計算する．積集合が空集合でなければ，左側のノードの左端と右側のノードの右端を範囲とするノードを新たに追加する．この処理を融合処理と呼ぶことにする．追加ノードには，先ほど計算した積集合が候補コード番号集合として対応づけられる．

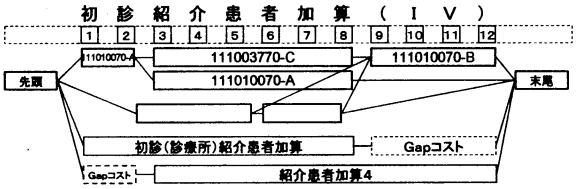

図 4：拡張中のラティス

以上のようなノードの連接処理を繰り返すことにより，候補コード番号が唯一になるノードや，元のマスターに対応するノードが出現してくる．それらのノードはお互い disjoint とは限らないため，各被覆間で優先順位をつけるため，“被覆のコスト”として，融合処理コスト，マスター再現性コスト，ノード間連接コストを定義した．

マスター再現性コスト

構成途中の部分文字列と、対応する候補マスター文字列との違いを表す尺度．前処理段階で，マスターは重要度の異なる複数の部分文字列に分解される．以下の式において，X，Y，Z はマスターの各パートの部分文字列長，X’，Y’，Z’は，ノード中に再現された各部分文字列の長さをあらわし，α，β，γは，あらかじめ決められた重み付けのための定数である．すなわち，マスターの文字列の，重み付け再現率を表す．

$$-\log((\alpha X'+\beta Y'+\gamma Z')/(\alpha X+\beta Y+\gamma Z))$$

融合処理コスト

コード番号の一致する部分文字列同士を結合する際のコスト

定数α´ × Gap 文字数

ノード間連接コスト

①先頭および末尾：

先頭とパス中の最左ノード(もしくは最右ノードと末尾）との間に存在するストップワード以外の文字列長÷入力文字列長

②パス上のマスター間：

ノード間に存在するストップワード以外の文字列数÷入力文字列長 ＋ マスター間連接コスト

各パスについて，被覆コストを以下のように定義した．コストの計算は，ラティスの拡張をする過程において，動的計画法の枠組みに従って行う．

$$PathCost(p,q)=\min_{p<t\le s\le q}\left\{\begin{matrix}PathCost(p,s)+\\OuterGapCost(s,t)+\\ConstCost(t,q)\end{matrix}\right\}$$

$$ConstCost(p,q)=$$

$$InnerGapCost(p,q)+InnerCompositionCost(p,q)$$

「マスター間連接コスト」について補足する．摘要欄の記述性質として，部分的な項目の並びに着目した場合，同じ種類の項目が並ぶ場合が多い．例えば「生化学的検査判断料」という項目が出現した直後には，次の点数回数，診療識別，アスタリスクが出現するまでは，「γ－ＧＴＰ」，「クレアチニン」，「ＧＯＴ」，「尿酸」等の実際の検査細目が連続する傾向がみられる．連接コストを考慮することにより，2.1.2における問題の一部が解消できる．

図 4の例では，入力の先頭から末尾までのパスはいくつも存在する．そのうちコスト最小のものを選択するが，ここでは，パス 1 とパス 2 を例としてコストの計算例を示す．

コストパス 1 = InnerCompostion(0, 8) + InnerGap(0.8) + OuterGap(9,12)

$= -\log((\alpha*3+\beta*0+\alpha*6)/(\alpha*3+\beta*3+\alpha*6)) + 2/12$

> 0.167

コストパス 2 = OuterGap(0,2) + InnerCompsotion(3, 12) + InnerGap(3, 12)

$= 2/12 + 0 + 0 = 0.167$

計算の結果，コストの低い「紹介患者加算４」を解として出力する．

4. 性能評価

4.1. 評価用データの分析

500 件レセプトから，人手で抽出された，約 18000 項目の摘要欄項目について正解文字認識データと正解コード化データを作成し，評価用データとして用いた．評価を行う前に，変換対象文字列がマスターに完全一致するもの，マスターに部分一致するもの，マスターを包含するもの，不一致の 4 種類に分類し，度数を調べた．約 18000 行のテキストから請求項目を抽出するのに，メモリ 120M，約 60 分を要した．

実行環境：Intel Celeron M processor 1.5GHz, 1.00G Ram,ノートパソコン

表 3：マスターデータの分類

データの個数 / マスター照合分類	
マスター照合分類	合計
不一致	7137
完全一致	5343
マスターを包含(複数解の場合あり)	5260
部分一致	254
総計	17994

4.2. 正解文字列のコード化精度評価

表 4に，画像の正解テキストデータを用いて，コード化した場合の精度について示す．再現率とは，本来レセプト画像中に存在し，コード化可能である文字列のうち，実際にコード化できたものの割合を表す．適合率は，システムが出力した解のうち，正解であったものの割合を表す（100 以下の場合，間違いを出力しているということを意味する）．まず，コード化対象データの構成について着目すると，変換対象文字列 17994 項目のうち，マスターと完全一致するものは，5343 件（100*5343/17994=29.7%）であり，これらについては，単純な辞書引き処理によっても正解可能なものである．提案手法により新たにコード化に成功したものは，合計 9752 件である．これにより，コード化率（再現率）を 84%にまで高めることができた．

表 4：テキストからコード化した場合の精度

データの個数 / マスター照合分類	コード化結果					
マスター照合分類	正解	偽陽性	偽陰性	結果総計	再現率	適合率
不一致	4646	1000	1764	7410	65%	82%
完全一致	5343			5343	100%	100%
マスターを包含(複数解の場合あり	5000	382	141	5523	1969%	93%
部分一致	106	32	116	254	2%	77%
結果総計	15095	1414	2021	18530	84%	91%

次に，同義定義ファイルへの依存度を検証するため，同義語定義ファイルなしでコード化した場合の精度を表 5に示す．

表 4結果と比べることで，同義語定義ファイルの貢献度もしくは依存度を判断できる．同義語定義 500 に対し，正解数は 13024 から 15095 へ増加している．

表 5：同義語定義なしでの精度評価

データの個数 / マスター照合分類	コード化結果					
マスター照合分類	0	1	2	結果総計	再現率	適合率
不一致	2601	1610	3125	7336	36%	62%
完全一致	5343			5343	100%	100%
マスターを包含(複数解の場合あ	4970	420	152	5542	94%	92%
部分一致	110	35	109	254	43%	76%
結果総計	13024	2065	3386	18475	72%	86%

4.3. 文字認識データのコード化精度評価

文字認識後に出力された摘要欄項目 7797 項目について，提案手法によるコード化精度の評価を行った．表 6に，文字認識結果データをコード化した場合のコード化評価結果をまとめた．縦軸は，文字認識結果の分類を表し，“A”は，文字認識が正しかったもの，”B”は，読み取ることができなかった（不読）もの，”C”は，謝った読み取りを行ったもの（誤読），”D”は，読み取っているが，文字抜け（脱字）があるものを表す．横軸はコード化結果の分離を表し，”1”,は，1 対 1 で正しく変換されたもの，”2”は，多対 1 で正しく変換されたものを表し，”4”,”5”は，いずれも誤って変換されたもの（偽陽性）を表す．また，”7”は，コード化されなかったもの（偽陰性），”8”は，コード化の必要がないものを表す．文字認識が正しい分類 A では，適合率 97%，再現率 88%と高い精度を示した．

表 6：文字認識後データのコード化精度評価

データの個数変換分類									
読取分類	1	2	4	5	7	8	総計	再現率	適合率
A	2116	223	61	17	242	170	2829	88%	97%
B	55	7	24	1	470	26	583	11%	71%
C	853	304	405	63	551	158	2334	53%	71%
D	1047	162	137	23	725	137	2231	58%	88%
総計	4071	696	627	104	1988	491	7977	64%	87%

本手法により，文字認識誤りを含むが，コード化に成功した事例を，表 7にいくつか示す．

表 7：本手法によりコード化に成功した事例

マスターコード	レセプト中の表記（文字認識後）	マスターの表記
642450095	/ケナコルトーＡ１関節腔内用皮内用]	関節腔内用皮内用ケナコルトーＡ１０ｍｇ
620000431	ルシドリール〔注射用〕　２５０？ｇ１瓶	注射用ルシドリール２５０ｍｇ
643310190	生食注［大塚１　２０ｍｌ？	大塚生食注２０ｍＬ
643310190	大塚生食注　１２０、？）　２Ａ	大塚生食注２０ｍＬ
120002270	？処方料特定疾患処方管理加？	特定疾患処方管理加算（処方料）
160093810	胃・十二指腸ファイバースコピー	ＥＦ－胃・十二指腸

4.4. 誤り事例の分析

同義語の登録不足以外の原因で，コード化の誤り原因としては，多い順に，“文脈不足”，“解重複”，“コストのチューニング不足”，“複数解”，“部分文字列”，などの原因が見られた．

文脈不足とは，文字だけでは判別できない例を表す．例えば，レセプトには「外来管理加算（月 3 回目まで）」としか書いてあるにもかかわらず，正解は「外来管理

加算（診療所）」が正解の場合である．これは，レセプト作成元の医療機関の種類が分かっていないと判断できない．解の重複は，本来正解が一つであるにもかかわらず，二重に出力してしまうものである．例えばレセプトに「C 反応性蛋白（CRP）定量」と書かれていた場合，括弧内の「CRP」は言い換えであるため本来無視して「CRP（定量）」のみ出力すればよいが，「C 反応性蛋白」と「(CRP) 定性」から，それぞれ「CRP（定量）」と「CRP」を出力してしまう．コストチューニング不足による誤り例は，「B-ALP」が典型的な例である．本来，血液中のアルカリフォスファターゼを表す「ALP」のみ出力すればよいのだが，被覆コストの関係から，「B-A」と「LP」を出力する．複数解とは，レセプトの記述が不十分で，解が一意に決められないものである．“文脈不足”によるものとの区別があいまいな場合もある．例えば「超音波検査（断層撮影法）」とだけかかれていると，「超音波検査(断層撮影法)(胸腹部)」と「超音波検査（断層撮影法）（その他）」のうち，何れなのか不明である．最後に，“部分文字列”と呼んでいるものは，マスターの文字列が完全に包含される場合に，偽陽性が出力される例を言う．例えばγ—ＧＴＰの，γが読み取れない場合，ＴＰが「総蛋白」として認識される．

4.5. 考察

現在の実装では，部分文字列を再配置する再に exact match を行っているため，文字認識間違いが比較的多く，候補統制用語のどの部分文字列も場合うまく機能していない．辞書の分割要素をテキスト上に配置する処理に，挿入，欠損，置換を許した k-difference inexact match[3] あるいは，置換のみを許す k-mismatch inexact match[3]を組み入れて，再現率の向上を図りたい．K-mismatch inexact match は，以下の事例のような，OCR 不読パターンに対して有効である．

例 A）外来？理加？　＝＞　外来管理加算

例 B）再？料（診療所）　＝＞　再診療（診療所）

例 C）処方料特定疾？処方？理加？　＝＞
特定疾患処方管理加算（処方料）

それぞれ，部分的に読み取ることに失敗しているが，読み取り文字数はあっている．

5. 関連研究

一般的に，二つの文字列間の編集距離を計算する方法として，DP マッチが知られており，テキストＴ中からパターンＰとの編集距離が k 以内の部分文字列を全て取り出す計算量Ｏ(kn)のアルゴリズムが知られている[3][4]．但し，本研究が課題とするタスクでは，数千，数万エントリーの辞書エントリーと近似計算をする必要があるため，DP マッチをそのまま適用するには現実的ではない．　テキスト Ｔ 中から複数パターンＰを同時に探索する問題は，exact set matching problem と呼ばれ，パターン集合の総長とテキスト長に対して，線形時間で解けることが知られている[4][5]．しかし，inexact set matching problem に関しては，あまり研究が進んでいない[6]．実装面では，パターン集合から作成した suffix tree の各ノードに DP マッチテーブルのカラムの一部を保持させる方法が知られている[6]．

しかし，これらの手法は，語順の違いを適切に扱うことができない．例えば，「大塚生食注２０mL」と「生食注（大塚）20mL」は，編集距離は大きいが，コード化にあたっては，同一視されなくてはならない．

6. まとめと課題

本研究では，OCR 読み取りエラーや用語の表記ゆれが存在する診療報酬請求書（レセプト）の記載文字列を統制用語にコード化する手法を考案し，有効性について検証した．本研究ではこの問題を，入力テキスト上に統制用語の部分文字列を最適配置する問題として捕らえ，動的計画法により，最小コスト配置を探索する方法を試した．

今後の課題：

1）精度向上（偽陽性の削減，再現率の向上）

- 分割要素に，k-mismatch inexact match を適用し，再現率の向上を図る．
- 分割単位を細かくする（形態素解析や n-gram の利用）

2）処理速度改善

- 無駄な探索空間の枝刈りを検討する．

文　献

[1] カルテ＆レセプト略語 8500，医学通信社，2004.

[2] 社会保険/老人保健 医科診療報酬点数表，社会保険研究所，東京，2004.

[3] G. Navarro, A Guided Tour to Approximate String Matching, ACM Computing Suveys,2001.

[4] Dan Gusfield, “Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology”, Cambridge Univ Pr (Sd),.1997.

[5] A. Acho and M. Corasick, Efficient string matching:an aid to bibliographic search, Comm.ACM, 18:333-40, 1975.

[6] G. Navarro, R. Baeza-Yates, and J.M. Arcoverde, Matchsimile:A flexible approximate matching tool for personal names searching, In Proceedings of the XVI Brazilian Symposium on Databases, pp228-242,,,,2001