様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２７　年　　６　月　　２　日現在

機関番号：　１２６０１
研究種目：　若手研究(B)
研究期間：　2012～2014
課題番号：　２４７００１３２

研究課題名（和文）医療文章からの部位表現の抽出と正規化

研究課題名（英文）Extraction and Normalization of Body site expression in Medical text

研究代表者
篠原　恵美子（山田恵美子）（Shinohara, Emiko）

東京大学・医学部附属病院・助教

研究者番号：４０５８２７５５

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：身体部位表現の内部構造モデルを構築し、身体部位表現を自動解析して内部構造を得る手法を提案した。これにより「角結膜」が角膜と結膜から構成されることを計算機で扱うことができるようになった。内部構造解析の結果は次に正規化をしなければならないが、解剖学用語と実際の身体部位表現を比較した結果、身体部位表現の正規化結果としてその要素になっている解剖学用語を抽出するのでは不十分であることが分かった。したがって正規化では解剖学知識の利用が必要である。そこで本邦で構築が進められている解剖オントロジーと身体部位表現を比較し、正規化技術の実現のために解剖オントロジーおよび自然言語処理技術に必要な事項を明らかにした。

研究成果の概要（英文）：I developed an internal structure model of body site expressions and algorithm to analyze an expression to get the internal structure. The expression is to be normalized based on the structure. I surveyed the body site expressions and anatomical terminology and the result shows that it proves inadequate to extract anatomical terms from the expression for normalization. Therefore, normalization requires the anatomical knowledge. I surveyed an anatomical ontology developed in Japan and body site expressions and revealed what normalization requires the ontology and the natural language processing technique.

研究分野：　医療情報学

キーワード：　自然言語処理　オントロジー　医学　用語

１．研究開始当初の背景

病院情報システムの普及とともに診療情報が電子データとして大量に蓄積されつつあり，その利活用に関心が集まっている．診療情報には検査値などの数値データやX線などの画像データなど様々な形式のものが含まれるが，中心となるのはテキストであり，患者の訴える症状やこれまでの経過，過去に罹った病気，診察で得られた所見，治療方針などについて記述される．ここから病名や検査値，部位表現，薬剤名などの要素情報を抽出することは診療テキストの利活用において基本的な要求である．本研究では先行研究で抽出精度が低いと報告された部位表現に着目し，それが実際に表す身体部位がどこなのかの特定を目指す．

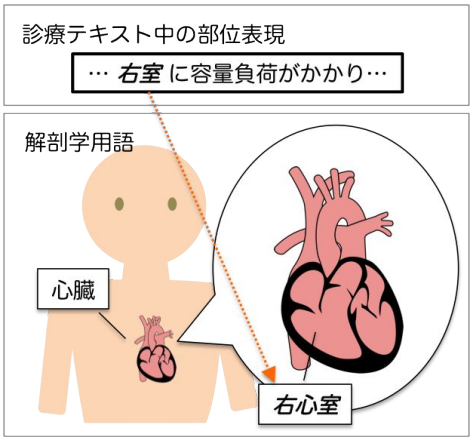

図 1. 身体部位表現と実際の身体位置

他の要素情報と比べ，部位表現の抽出が難しいのは，文脈のバリエーションに起因する．部位表現は単独で出現するだけでなく，「膝関節痛」のような症状表現，「狭心症」のような病名の一部としても出現する．どのような場合を抽出対象とするかを明確にしておくことが精度の高い抽出を実現するのに必要である．

抽出された部位表現からそれが指し示す実際の身体部位を特定する際には，部位表現そのもののバリエーションが問題となる．

身体部位を記述するための語は解剖学用語であり，既に学会等から複数の用語集が編纂・出版されており，さらには用語間の関係や意味を記述したオントロジーの構築も進められている（以降，標準部位表現）．標準部位表現は身体部位と紐付けられているため，「実際の身体部位の特定」は部位表現を標準部位表現に関連付けることで実現可能である．ところが，診療テキストで用いられる部位表現（以降，カルテ部位表現）は標準部位表現だけでなく，表 1 に示す 4 つのバリエーションを含む．

【標準部位表現の異表記】
標準部位表現と（ほぼ）同じ概念を表す部位表現で，標準部位表現とは異なる表記である．異表記には①同義語，②省略などがある．

【標準部位表現と異なる粒度の部位表現】
粒度の異なりには高低両方がある．粒度の低い表現は複数の標準部位表現が示す部位全体を表しており，標準部位表現の③上位語である．粒度の高い表現は標準部位表現が示す部位の一部を表しており，標準部位表現の④下位語である．

表 1. 身体部位表現のバリエーション

type		example	
		カルテ部位表現	対応する標準部位表現
表記	①同義語	尿路	尿管
	②省略	腎	腎臓
		S状間膜	S状結腸間膜
粒度	③上位語	上部消化管	食道，胃，十二指腸
		角結膜	角膜，結膜
	④下位語	右股関節	股関節
		大腿裏	大腿

これらのうち，①同義語と③上位語の一部については数が限られており網羅的な収集が可能であると考えられる．一方，②省略と③上位語の一部，および下位語は生産的であり，網羅的な収集は困難である．このような生産的なカルテ部位表現については，図 2 のような内部構造を利用することで，標準部位表現と関連付けることが可能と考えられる．

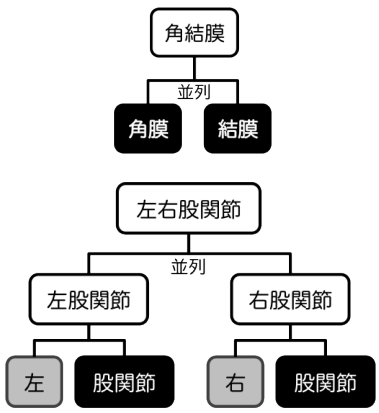

図 2. 内部構造

２．研究の目的

本研究では診療テキストで用いられるカルテ部位表現を抽出し，標準部位表現と関連付ける，即ちカルテ部位表現が示す実際の身体部位を特定する技術の研究・開発を行う．

３．研究の方法

(1) 正規部位表現モデルの構築

身体部位は標準部位表現と修飾語の階層的組み合わせにより表現可能と考えられる．本研究ではこのような枠組みに基づく表現を「正規部位表現」と呼ぶ．本段階では正規部位表現の枠組み，即ち「正規部位表現モデル」を構築した．

具体的には，言語学分野における語形成の知見を援用し，これを解剖学用語に適用した．そのためには語形成の議論では十分に扱われていない，要素語の意味分類について検討する必要がある．部位表現は人体の中の位置を表すものであり，基準となる場所の識別子と，そのスコープ内での場所の識別子の 2 つから構成されると考えられる．この前提に立って解剖学用語を分析し，部位表現の要素語の意味分類を構築した．

この意味分類に基づいた語形成過程を再現するため，2 つの分類を合成した時に作られる語の分類を定める合成規則を作成した．

内部構造の要素の最小単位は文字とした．「ラムダ」などカタカナ語の場合はこれを最小単位とした．これは図 2 の「角結膜」のように語を分解して合成するような場合を表現するためである．このような最小単位に上述の意味分類を付与した辞書を作成した．

表 2. 部位表現の要素語分類

[接辞]	[接頭辞]	[接頭辞-数字]
		[接頭辞-その他]
	[接尾辞]	[接尾辞-形状]
		[接尾辞-領域]
		[接尾辞-事態]
		[接尾辞-その他]
[内容語]	[持続物]	[生命体]
		[統合体]
		[非統合体]
	[特徴]	
	[構造]	
	[関係]	
	[従属的持続物]	[数字]
		[事態]
		[その他]

表 3. 部位表現の合成規則（抜粋）

前置要素語	後置要素語	合成語
[生命体]	[統合体]	[生命体]
[生命体]	[特徴]	[生命体]
[生命体]	[構造]	[生命体]
[生命体]	[関係]	[生命体]
[関係]	[生命体]	[生命体]
[統合体]	[統合体]	[統合体]
[統合体]	[特徴]	[統合体]
[統合体]	[構造]	[統合体]
[統合体]	[関係]	[統合体]
[事態]	[構造]	[統合体]

(2) 部位表現に含まれる解剖学用語と部位表現の位置関係

同義語を含めた標準部位表現を十分に整備しておけば，カルテ部位表現の内部構造には標準部位表現が現れるはずである．しかしこの標準部位表現をそのまま正規化結果として扱えない場合がある．例えば「食道周囲」に含まれる標準部位表現「食道」であるが，「食道周囲」は食道そのものやその一部ではない．そこで，部位表現に含まれる解剖学用語と，正規化結果とするべき解剖学用語との関係を調査した．

(3) 部位表現と解剖オントロジーの関係

上述した通り，内部構造に現れる標準部位表現をそのまま正規化結果として扱うことはできず，正規化のためには解剖学知識を使った推論が必要になると考えられる．また，正規化結果を十分に活用するためには，その正規化結果もまた計算機において解剖学知識体系中に位置づけられる必要がある．計算機上での解剖学知識はオントロジーとして整備が進んでおり，これに対して部位表現やその要素語を対応づけることが必要になると考えらえる．

その実現のための基礎とするため，部位表現と現状のオントロジー中の概念を人手で対応付け，①情報欠落が起こるか，②対応概念の数を調査した．②を調査対象としたのは，対応概念が複数あれば対応付けの技術的な難易度が高くなると考えたためである．

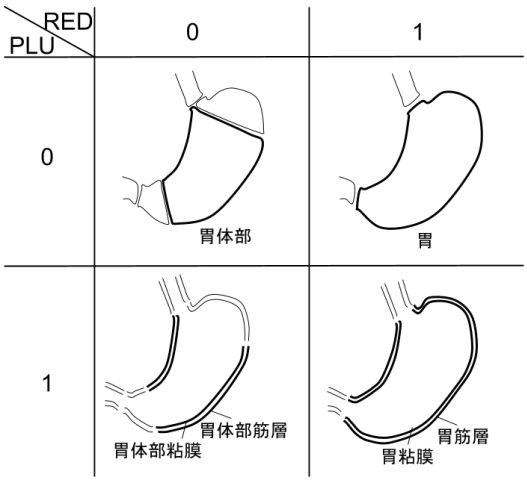

図 3. 「胃体部」とオントロジーの関係

各セルが 1 つのオントロジーに対応しており，セル中に示した用語は該当オントロジーに定義された概念を表す．また RED は情報欠落の有無，PLU は対応概念の数を表す．例えば右上のオントロジーでは「胃」のみが定義されており，「胃体部」の対応概念は「胃」のみである（PLU=0）．この場合，胃底部や噴門部でないという情報が欠落している（RED=1）．

４．研究成果

(1) 正規部位表現モデル

表 2 に要素語の意味分類，表 3 に合成規則の抜粋を示す．

モデルの初期評価として，解剖学用語 80 語に対して提案モデルに基づいた内部構造解析を適用したところ，入力と同じ語が形成可能であったのが 77 語，正しい合成過程を少なくとも 1 つ含んでいたものが 73 語，明らかな誤りでない合成過程を全て含んでいたものが 68 語であった．また，要素語に解剖学用語を多く含む内部構造を残すような枝仮アルゴリズムを適用したところ，61 語で適切な枝狩りが行われた．誤り事例を分析したところ，合成規則や辞書の整備で解決可能なものの他，要素語自体が省略表現であるために意味分類を誤った場合があった（例えば「皮質核線維」における「皮質」は「大脳皮質」の，「核」は「運動核」の省略である）．この省略の扱いは今後の課題である．

(2) 部位表現に含まれる解剖学用語と部位表現の位置関係

調査は症例報告で用いられた部位表現 256 種を対象として行った．調査の結果，部位表現のうち 34.9%は単独の解剖学用語，30.6%は複数の解剖学用語，27.1%は解剖学用語とそれ以外の語の組み合わせ，7.5%は解剖学用語を含まないものであった．また，27.0%の部位表現は正規化結果となる語が含まれていないものであり，「骨盤内」など含まれている用語と隣接する領域や範囲を示す場合や，「(眼圧は）右（19㎜Hg)」など相対位置による表現があった．

この調査結果から，部位表現の正規化には，内部構造解析のみならず，隣接領域等を推論するための医学知識が必要であることが分かった．

(3) 部位表現と解剖オントロジーの関係

調査は症例報告で用いられた部位表現 154 種および 2014 年 12 月時点の解剖オントロジー（概念数 198,377）を対象として行った．調査の結果，情報欠落が起きたものが 29，複数概念に対応づいたものが 42 あった．事例を分析したところ，オントロジーに追加するべき概念として，「右示指」のような領域を表す用語，「中肺野」のように解剖学用語ではないが臨床で使われる用語，また「内側」「肛門側」など各概念における方向概念があった．

また，情報欠落が無く対応概念が 1 つである事例のうち，部位表現とオントロジーに記述されている概念の名称が異なるものを分析した．その結果，「中手指節関節」と「MP 関節」のような同義語や「腎」と「腎臓」のような軽微な表記ゆれの他に，「胃底腺」と「胃底部」の一部である「固有胃腺」のように部位表現の内部構造と解剖学知識から同義と判断可能なもの（図 4）があった．同義

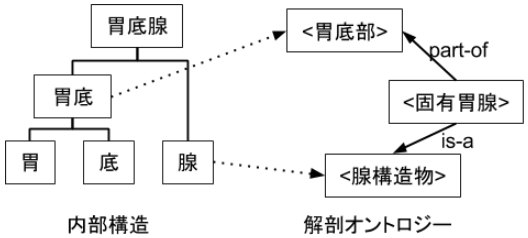

図 4.内部構造とオントロジーによる正規化
まず内部構造の要素語である胃底と腺を正規化する．”胃底腺”の内部構造では胃底が腺を修飾しているので，解剖オントロジーで<胃底部>に関連を持つ<腺構造物>を探すと<固有胃線>が得られる．

語辞書の整備，および内部構造やオントロジーを利用した推論機構の実装が必要であることが分かった．

５．主な発表論文等

（研究代表者，研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 2 件）

[1]. 篠原恵美子．高精度な正規化技術の実現に向けた身体部位表現の内部構造モデル．医療情報学（査読有）34(5), 2014, 211-20.
[2]. 篠原恵美子，今井健，大江和彦．身体部位表現の正規化における処理スキームの提案．医療情報学 (査読有) 34(Suppl.), 2014, 322-5.

〔学会発表〕（計 1 件）

[1]. 篠原恵美子，今井健，大江和彦．身体部位表現と解剖オントロジーのマッピングに関する基礎的検討．第 19 回日本医療情報学会春季学術大会，2015 年 6 月 13 日，仙台国際センター（宮城・仙台）．

６．研究組織

(1)研究代表者

篠原　恵美子（SHINOHARA, Emiko）
東京大学・医学部附属病院・特任助教
研究者番号：40582755