文章構造化技術の医用画像データベースへの応用

Application of Text Structuring Technologies to a Medical Picture Database

安永　晋＊　川上洋一＊　笹井浩介＊
Anei, Shin　Kawakami, Youichi　Sasai, Kosuke

要旨

電子化された大量の情報から必要な情報を効率よく抽出することが要求されている。そこで我々はユーザーが入力する情報から得られた知識や経験を用いて学習する新しい概念の情報検索技術を開発している。本技術のイメージング分野への応用としてPACS（Picture Archiving and Commuincating System: 医用画像保管・転送システム）と連携した医療診断支援システムがある。我々は医用画像及び100文字程度の診断レポートから必要な情報を抽出することにより症例を蓄積し、診断支援に利用する知的なシステムの実現を目指している。この目的に向け、機能評価のためのソフトウェアを開発した。本論文ではこのシステムに要求される言語処理技術について開発したアルゴリズムを中心に述べる。

Abstract

The efficient extraction of information from large electronic databases is essential to a practical database system. With this in mind, we are developing a novel data search system which automatically tracks the particular uses to which individual users put the system. Applied to the imaging domain, we are constructing an intelligent aided diagnosis system which can store the pertinent information extracted from medical images and short (approximately 100-character) diagnosis reports, and which can be integrated into a PACS (picture archiving and communicating system). This paper focuses on the system's language processing algorithm.

1　はじめに

電子化された大量の情報を有効に活用するには、ユーザーの必要としている情報が適切に検索可能であることが必要である。しかし、従来のデータベースシステムでは、情報を発信する側の考え方で情報を作成し蓄積しているため、ユーザーが適切に情報を得ることが難しかった。

この解決のため、我々はユーザーが入力するいろいろな情報からユーザーの知識や経験を吸収し、それを利用してデータ構造と検索論理をアクティブに進化させることが可能な情報検索技術の開発に取り組んでいる[1)]。このような情報検索が実現できると、ユーザーの意図を理解して適切な情報を提供するシステムが実現できる。

このシステムは広い汎用性を有しているが、そのひとつの応用例として、PACSと連携した医療診断支援システムがある。これまでは過去の症例が医療診断支援に適した構造で蓄積されていなかったので、医療現場に対して診断支援のための情報をリアルタイムに提供することは困難であった。しかし我々の開発しているシステムでは過去の症例を医療現場が再利用しやすいデータ構造で蓄積し、それを医療現場の意図を理解してリアルタイムに提示することが可能になる。

本論文では、まずユーザーの知識や経験を利用してデータ構造と検索論理を進化させる処理について述べた後、医療診断支援システムを実現するためのコア技術の1つである言語処理技術について述べる。

2　データ構造と検索論理の進化

ユーザーの知識や経験を反映させ、アクティブにデータ構造と検索論理を進化させるための情報処理フローを以下に示す。

⑴XML（eXtensible Markup Language）を交換構文としてタグ付けされた情報をデータエレメントとする。
⑵RDF（Resource Description Framework）スキーマによりデータエレメントの関係を構造化する。
⑶利用者から提供される情報をアクティブに利用してデータ構造と検索論理を動的に進化させる。

XMLとは、データ交換を目的としてWWWコンソーシアムが標準勧告しているマークアップ言語であり、これを用いることによってテキストデータに限らず様々なデータに対してデータの意味を同時に記述することができる。

RDFとは、WWWコンソーシアムで1999年2月に正式に勧告された最先端のスキーマで、簡単なルールであたかも人間の脳の情報処理のようにデータエレメントとその関係を動的に記述できる。

＊コニカミノルタテクノロジーセンター㈱　システム技術研究所
イメージシステム開発室

例えば「論文：情報処理の高速化の著者は田中太郎である」という情報に対し、RDFでは「論文：情報処理の高速化」を主語（リソース）、「著者」を述語（プロパティ）、「田中太郎」を目的語（リテラル）と呼び、主語、述語、目的語の三つを一組として情報を**Fig. 1** のように表現する。

複雑な情報の場合、目的語の部分に別のリソースを指定することにより**Fig. 2** のように表現する。その結果、データエレメントとその関係が「数珠つながり」になる。

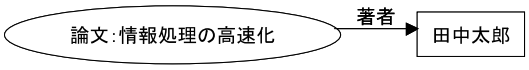

Fig.1 RDF data structure

Fig.2 RDF data structure of complicated information

このスキーマの利用によりデータ構造を柔軟に変化させることができ、新たなデータやユーザーから知識を吸収してデータおよびデータ構造をアクティブに進化させることができる。またその時々に抽出ルールを入力すれば、すでに構造化された関係を利用して新たな情報が抽出できる。

本システムにおける情報処理フローを**Fig. 3** に示す。

Data sourceから得られたデータはTranslatorsにより解析され、XMLを交換構文としたRDFで構造化されてData Storeに保存される。

ユーザーによって入力された質問は、Execute compositionによって解析され、それをもとにQuery engine・Miningが検索処理を実行する。その際には必要に応じて「知識・構造・発見モデル」を利用して個々のユーザーに対して提供する情報の最適化や効率の良い質問を導出するためのヒントを提示する。

このシステムが利用される過程で、ユーザーの知識や経験から得られた情報はRealization storeにおいて「知識・構造・発見モデル」に追加される。そして、Restructure engineは情報が追加された「知識・構造・発見モデル」を利用してデータの構造をアクティブに進化させる。

この繰り返しにより、システムが使い込まれていくほど「知識・構造・発見モデル」に情報が蓄積され、システムが進化していく。

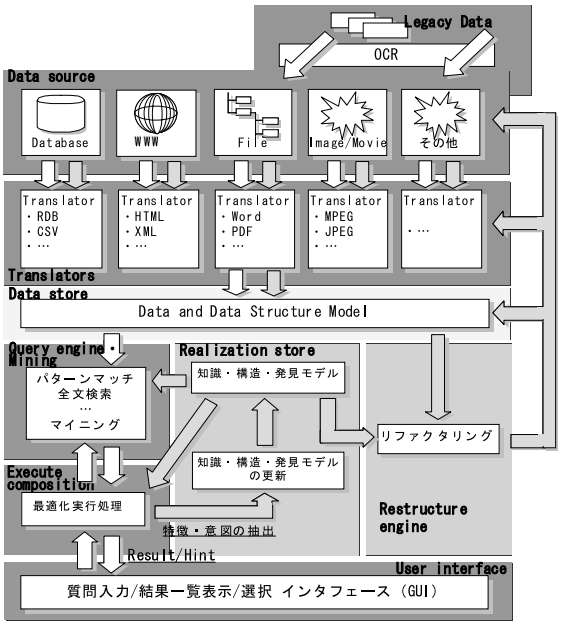

Fig.3 Information processing flowchart

3　医療診断支援システムへの応用

読影医は、CTやMRIなどで撮影された大量の放射線画像から特徴的な画像を選び、その特徴を100字程度の短文（以下「レポート」と呼ぶ）で記録する。その際に過去の症例が的確に提示されれば、効率良く診断を行う上で好都合である。我々が検討しているシステムでは、レポートの内容を画像と関連付けて構造化し、それを症例データベースとして利用することによって診断支援を可能にする。さらに、診断を行うごとにその結果が新たな症例として追加され、これによって動的に症例データベースの内容が強化される。

3．1　診断支援の手順

まず、第1のステップとして、レポートから**Fig. 4** に示すように必要なキーワードを抽出し構造化を行う。具体的には、まず形態素解析を行ってレポートを単語に分解し、得られた単語の「属性」（部位なのか症状なのか診断なのか、など）を判定し、この判定をもとにキーワードを抽出する。単語の属性の判定を行うためにはシソーラス（階層構造の形になっている類義語辞書）を用いる。本システムでは、シソーラスとして、医学中央雑誌刊行会が発行する「医学用語シソーラス」を用いてい

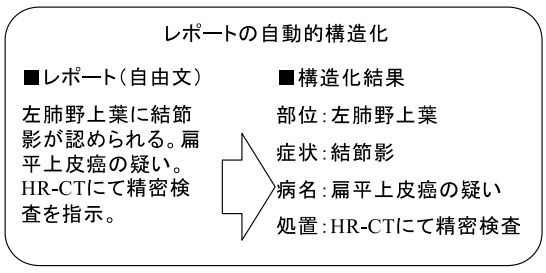

Fig.4 Structuring of the report

る。キーワードを正しく抽出できなかった場合は、ユーザーがキーワードを修正することも可能である。

第2のステップとして、第1のステップで得られたキーワードをキーにして、ユーザーの意図に沿った優先順序で情報を提示するための診断支援モデルを構築する。診断支援モデルの詳細は3．2で述べる。

第3のステップとして、読影医が第2のステップで得られた診断支援モデルを参考にして診断を記入し、「放射線画像・元のレポート・キーワード・診断」の組を症例データベースに保存する。この時、ユーザーがキーワードを修正して保存した場合、修正されたキーワードをシソーラスに追加する。

症例データベースおよびシソーラスは、症例が増えるごとにデータ量が大きくなっていく。よって使われれば使われるほどこの診断支援システムは成長していく。

3．2　診断支援モデル

症例データベースには、あらかじめ過去の症例における「放射線画像・レポート・キーワード・診断」の組がRDFを用いて関連付けられた形で保存されている。

この症例データベース内を第1のステップで得られたキーワードをキーにして検索し、得られた症例をシソーラスの情報をもとにRDFを用いて関連付けて表現したものが診断支援モデルである。一例を**Fig. 5** に示す。

例えば「部位」に対応するキーワードを検索キーとすれば、その部位の上位・下位の概念やその部位に関連付けられている「症状」や「診断結果」などを得ることができる。検索キーに「症状」など他の情報も用いれば、さらに結果を絞り込むこともできる。

RDFを利用してすべてのデータを的確に構造化することにより、部位や病変などが明らかになればそこから導かれる過去の画像データを含む症例を提示できる。これにより診断支援を行うことができる。

4　「レポート構造化」に用いる言語処理技術

ここでは、3で述べたシステムの「レポート構造化」に用いる言語処理技術の詳細を述べる。

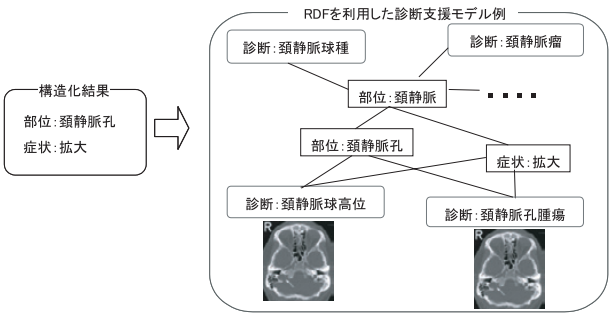

Fig.5 Diagnostic support model

4．1　キーワードの抽出

Fig. 6 に処理フローを示す。これに沿って説明する。

まず、入力した文章（レポート等）に対して形態素解析の処理を行う（ステップ1）。

形態素解析には、奈良先端技術大学院大学の松本研究室が開発した日本語形態素解析システム「茶筌」（http://chasen.naist.jp/）を用いている。

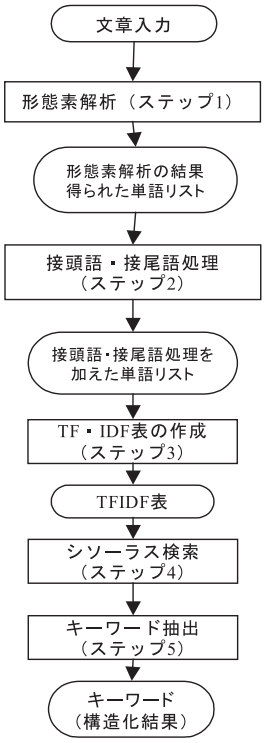

Fig.6 Keyword extraction processing flowchart

Fig. 7 に形態素解析処理の一例を示す。

形態素解析には辞書（単語のリスト）が必要であり、「茶筌」にもあらかじめ辞書が付属している。辞書に含まれない語は形態素解析で抽出されることはない。辞書に新しく語を追加することにより形態素解析の精度が上がっていく。

本システムでは、シソーラスに含まれている単語はすべて辞書に追加している。

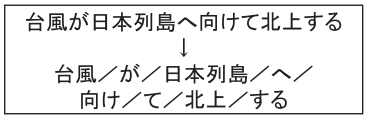

Fig.7 Morphological analysis processing

次に、接頭語・接尾語の処理を行う（ステップ2）。

これは、「左」「右」「上」「下」などの「接頭語」や「内」「部」などの「接尾語」の候補リストをあらかじめ用意しておき、これらの語が名詞と連続している場合は1つの語として扱う処理である。

例えば、「頚動脈」という単語は形態素解析用の辞書にあるが、「左頚動脈」という語は辞書にないとする。この場合、形態素解析では「左頚動脈」は「左」と「頚動脈」に分解されるが、これを「左頚動脈」一語として扱う。

次に、形態素解析で得られた単語のリストから、各単語の「TF・IDF値」を計算し、単語とTF・IDF値の組のリスト（以下、TF・IDF表と呼ぶ）を作成する（ステップ３）。

まず、TF・IDF値について簡単に説明する[2)]。

TF値は、その文章中に単語が出現する回数を表す。

IDF値は、「コーパス（サンプルとして集めた文書の集合）中でその単語が現れる文書の割合」を表し、その単語が現れる文書の割合が高いほど「IDF値」は小さくなる。

このTF値とIDF値の積がTF・IDF値である。

一般に、当該文書中に出現する頻度が高く、なおかつ他の文書にはあまり出現しない語ほど、当該文書を特徴づける度合いが高い。従って、TF・IDF値が大きい単語ほどその文書を特徴づける度合いが高いと言える。

次に、作成したTF・IDF表にある各単語についてシソーラスを検索して属性を求める（ステップ４）。

シソーラスの構造の一例を**Fig. 8** に示す。

シソーラスに含まれる各単語には構造上の位置を示す数字（シソーラスコード）が与えられている。

TF・IDF表にある各単語についてシソーラスを検索し、得られたシソーラスコードを手がかりに属性を求める。ステップ２の処理で接頭語・接尾語がつけられた単語については、この接頭語・接尾語は除いて検索する。

例えば、シソーラスにおいて、大分類として「部位」という分類があり、これを表すシソーラスコードが「A」であったとする。この時に「部位」という属性を考える場合、「A」で始まるシソーラスコードを持つ語が「部位」という属性を持つと判断する。１つの語が複数のシソーラスコードを持つ場合もあるので、１つの語が複数の属性を持つ場合もありうる。

単語がシソーラスに見つからなかった場合は、その単語の属性は不明ということになる。

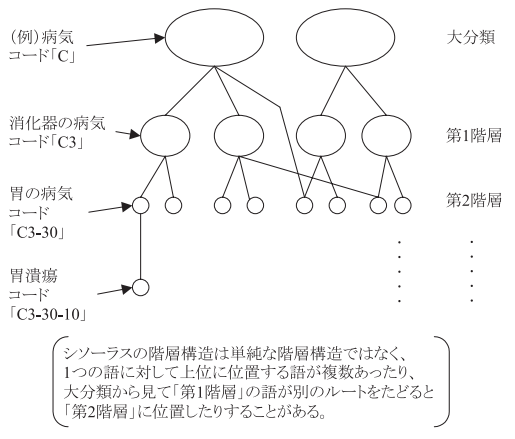

Fig.8 Thesaurus structure

そして、何種類かの属性について、抽出されたその属性を持つ単語の中で最も「TF・IDF値」の大きい単語を、入力した文章のその属性に対応するキーワードとして抽出する（ステップ５）。もちろん、属性に対応するキーワードが存在しないこともありうる。

具体的な文章の例を挙げる。

「大腸にポリープが多発している。特にS状結腸に密集している。」という文をこの方法で処理する場合を考える。

属性として「部位」「症状」の２つを考える場合、「部位」に対応するキーワードの候補として「大腸」「S状結腸」の２つが、「症状」に対応するキーワードの候補として「ポリープ」が抽出できる。

「部位」に対応するキーワードとして「大腸」「S状結腸」のどちらを選択するかを「TF・IDF値」で判定すると、「大腸」は「S状結腸」に比べてよく使われる用語なので、一般的には「IDF値」は「S状結腸」の方が大きくなる。コーパスの選び方によってはそうでないこともあるので、コーパスの選択は適切に行う必要がある。「TF値」はどちらも同じ（１回だけ出現している）なので、「S状結腸」の方が「部位」に対応するキーワードとして選択される。

ここで抽出されたキーワードは、３で述べたシステムの「診断支援モデル」の検索キーとして用いられ、またレポートをデータベースに保存する際に「診断支援モデル」の一部となる。抽出されたキーワードが適切ではないとユーザーが判断した場合は、ユーザーが修正することもできる。

４．２　新しい単語の追加

前述したように、形態素解析用の辞書に含まれない語はキーワードとして抽出できない。また、辞書に含まれていてもシソーラスに含まれていなければその単語はキーワードとして抽出できない。

しかし、これでは新しく生まれた用語などはキーワードとして抽出することは不可能なので、単語を新しく辞書やシソーラスに追加できるようにする必要がある。

辞書への単語の追加については、単純に追加したい単語を既存の辞書に付け加えればよい。一方、シソーラスに登録するには構造上のどの位置に追加するか（すなわち、シソーラスコードをどう与えるか）を決めなくてはならない。

ここでは、新しく追加したい単語について、このシソーラスコードを推定する方法について述べる。

このシステムでは、３で述べた第３のステップにおいて、データベースに「放射線画像・元のレポート・キーワード・診断」の組を登録する際に、キーワードをシソーラスおよび辞書に追加する処理を行う。実際にシソーラスや辞書に単語が追加される可能性があるのは、ユーザーが抽出されたキーワードに修正を加えた場合で

ある。なぜなら、本来キーワードはシソーラスを用いて抽出されているからである。

Fig. 9 に単語を新しくシソーラスに登録する場合の処理フローを示す。単語をシソーラスに登録する場合は「属性」との組で登録する。上述のキーワードをシソーラスに登録する処理では、「属性」としてそのキーワードに対応している属性を指定する。

まず、登録したい単語を形態素解析する。以下の処理は、形態素解析によって複数の単語に分解された場合と分解されなかった場合で異なる。

複数の単語に分解された場合は、分解されたそれぞれの単語をシソーラスから検索する。そして、その中で1つでもシソーラスに含まれる語がある場合、もとの単語の構造上の位置は分かれた単語の「すぐ下」にあると推定する。つまり、分解された単語のシソーラスコードが「A1-20-30」だった場合、もとの単語のシソーラスコードは「A1-20-30-990」のように推定する。分解された単語が複数のシソーラスコードを持つ場合は、もとの単語も複数のシソーラスコードを持つと推定することになる。シソーラスに含まれる語が複数あった場合は、後方の語の「すぐ下」にあると推定する。

なお、分解された語の中にシソーラスに含まれる語がなかった場合は、この方法では推定できない。この場合は「推定不能」と判定する。

次に、推定したシソーラスコードが「属性」と合致するかどうかを確認する。具体的には、4のステップ4で述べた方法で、推定したシソーラスコードに対応する「属性」を求め、これが入力した「属性」と合致するシソーラスコードのみを単語に与える。

この結果、単語にシソーラスコードを全く与えられなかった場合、及びシソーラスコードが「推定不能」の場合は、「属性」に対応するシソーラスコードを1つ生成してそれを単語に与える。上と同様の例で「部位」という属性を考えた場合、「A」で始まるシソーラスコードが対応するので、例えば「A99-990」というコードを与える。

登録したい単語を形態素解析した結果、複数の語に分かれなかった場合は、登録したい単語をシソーラスから検索する。登録したい単語が見つかった場合は、そのコードが「属性」に反するかどうかをチェックする。反しない場合は特に何もする必要はない。反する場合は、「属性」に対応するシソーラスコードを1つ生成してそれを追加する。登録したい単語が見つからなかった場合は、「属性」に対応するシソーラスコードを1つ生成してそれを単語に与える。

具体例を挙げる。

「胃潰瘍」という単語がシソーラスに登録されている場合に、「神経性胃潰瘍」という語を「病名」という属性で追加したいとする。「神経性胃潰瘍」を形態素解析すると「神経性／胃潰瘍」と分解される。「胃潰瘍」のシソーラスコードを「C5-60-70」とすれば、「神経性胃潰瘍」のシソーラスコードは「C5-60-70-990」と推定される。「病名」という属性が「C」で始まるシソーラス

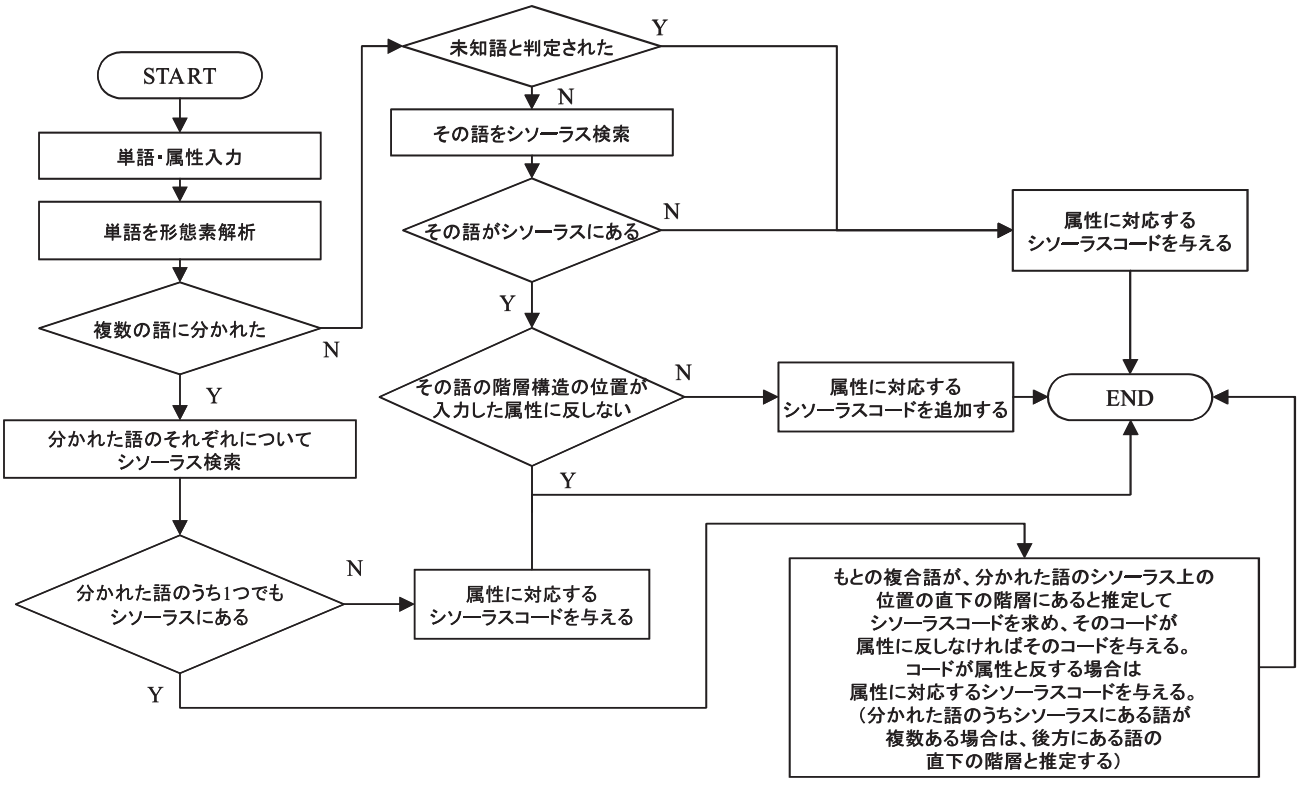

Fig.9 Flowchart of process registering a word into the thesaurus

コード対応しているとすれば、この推定は属性に反しないので、推定した「C5-60-70-990」というシソーラスコードを与える。もし「病名」ではない別の属性、例えば「部位」という属性でこの単語を追加しようとした場合は、この推定は属性に反している。よってこの場合は、属性に対応するシソーラスコード、例えば「A99-990」というコードを生成して与える。

5　現状の性能と今後の課題

5．1　現状の性能

一般的に入手できる言語処理ツールでは読影レポートに使用される専門用語の属性判別はほとんど不可能である。しかし本報告で述べてきたアルゴリズムと辞書およびシソーラスを組み合わせて実際の読影レポートからキーワード抽出を行った結果、50%～70%の確率で適切なキーワードが抽出できた。また実際の医療機関で使い込めばユーザーの知識や経験を吸収することによって一層の精度向上が期待できる。しかし初期性能と汎用性を向上させるためには新たな課題があることも明らかになった。

なお評価には病院などで公開されており一般に入手可能な読影レポートを使用した。

5．2　今後の課題

現在は4．2で述べた通り、新しい単語を登録する場合は、まず新しい単語を分解し、分解された語のシソーラスコードからもとの単語のシソーラスコードを推定している。難点は、分解された語の中にシソーラスに含まれる語がない場合にはこの方法が使えないという点である。そこで、登録したい単語は文章中で他のどの単語と同時に使われているか、という情報を用いてシソーラスコードを推定する方法を検討する。

一方、例えば「異常は見られない」という文からキーワードを抽出する場合を考える。「異常」という単語がシソーラスにある場合は「異常」がキーワードとして抽出されることになるが、「異常は見られない」という文のキーワードが、「異常」というのは適切とは言えない。そこで利用を考えているのが「係り受け情報」である。「係り受け」とは、文を文節単位に区切った場合に、文節と文節のつながりのことである。これを使うことによって、例えば「ない」という単語を含む文節に係っている文節にある単語は抽出しない、といった処理方法などが考えられる。

また、4．1の処理においてキーワードを抽出する際、現在はTF・IDF値が最大のものを属性ごとに求めて抽出している。例えば「心臓が肥大し、動脈も硬化している」という文に対し、属性として「部位」と「症状」を考えると、「部位」に対応するキーワードの候補として「心臓」「動脈」の2つが、「症状」に対応するキーワードの候補として「肥大」「硬化」の2つがある。この時にTF・IDF値によって「心臓」と「硬化」を抽出してしまうと、もとの文と全く異なる内容になってしまう。

こういった誤りを避ける処理にも、上述した「係り受け」が利用できると思われる。上の例だと、まず「心臓」を抽出した後、「心臓」を含む文節「心臓が」は「肥大し」に係っているので、「肥大」を優先的にキーワードとして抽出するといった方法が考えられる。

6　まとめ

PACSと連携した医療診断支援システムにおいて、放射線読影レポートに言語処理を行った上で画像とともに構造化するシステムを開発した。これにより、画像及びレポートを再利用し、診断支援を有効に行うことができる。現時点では機能評価のためのソフトウェアを開発した段階であるが、今後はこのソフトウェアを専門の先生方に評価していただき、システムとしての性能のさらなる向上を目指す。

謝辞

本研究を実施するにあたって日本医療情報学会に加入されている以下の先生方に多大な協力をいただきました。ここに感謝の意を示します。

京都大学医学部付属病院　黒田講師
京都大学医学部付属病院　竹村助手
大阪大学歯学部　玉川助教授
関西医科大学　仲野講師
大阪市立大学医学部　朴助教授
兵庫医科大学　平松講師
大阪大学医学部　松村助教授
兵庫医科大学　宮本教授
（50音順）

●参考文献

1）笹井浩介：「利用者の意図が理解できるデータベース検索システムの開発」，月刊ファームステージ9月号，技術情報協会（2004）
2）徳永健伸：「言語と計算- 5 情報検索と言語処理」，辻井潤一編，東京大学出版会（1999）