医療情報のフィルタリングと自動タグ付け

1L−2

寿恵村唯子、嶋本公徳、岩下達哉（新日鉄ソリューションズ）、
村山　博、大貝晴俊（新日本製鉄）、中島律子（科学技術振興事業団）、
野村浩郷、中村貞吾、永井秀利、井上大悟、坂田大輔（九州工業大学大学院　情報工学研究科）

1．はじめに

インターネットにおける信頼性・高速性・利便性・安全性に関する問題の解決に資するため、遠隔地重粒子線がん照射影響シミュレータを業務として想定した「高度医療ネットワークに関する調査研究」が、平成 10 年度から科学技術振興調整費に基づいて進められている。科学技術振興事業団では、その中で「情報の収集技術の研究」として医療テキスト情報の収集とファクトデータベースの構築研究を担当した。本報告では、その内容について述べる。

2．医療情報収集研究の目的と医療情報収集システム構成

2.1 研究の目的　医療機関の持っている各種画像、検査データ、学術論文等の情報を自動収集し、ファクト情報およびリンク情報を自動抽出するシステムについて研究し、病名や診療方法などの各要素からなる医療ファクトデータベースを構築し、その応用について研究する。

2.2 医療情報収集システム構成

図１に示すように医療情報収集、医療ファクトデータ抽出、医療情報加工の３つのサブシステムより構成される。

(1) 医療情報収集システム
特定機関医療情報収集機能と広域医療情報収集機能より構成される。

(2) 医療ファクトデータ抽出システム
医学用語辞書作成機能、タグ付け機能、テンプレート生成機能、ファクト抽出機能、データベース格納機能から構成される。

(3) 医療情報加工システム
データベースからの指定情報の抽出可視化機能、類似論文検索機能等から構成される。

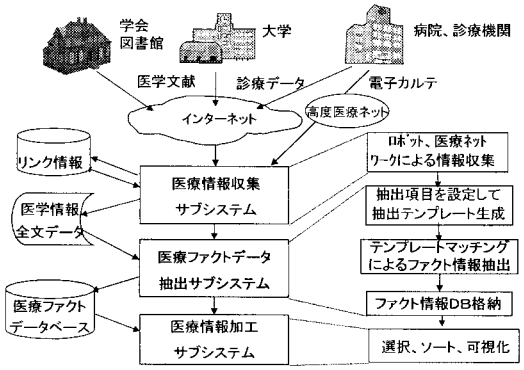

図１　医療情報収集システムの概要

3．インターネットからの医療情報収集

広域医療情報収集は、ロボットの収集した URL を医療情報フィルターで選別している。

3.1 収集ロボット

国立がんセンターを起点に、ロボットで URL を収集。

3.2 医療情報フィルター　各 URL の情報に含まれる医学専門用語の個数、論文の表現用語の個数を用いて、①医学情報、②医学論文を選別している。画像を含んでいるかどうかも判定している。

判定ロジック

1) 医学専門用語種別ごとの個数を算出
　治療方法、医療薬、診断方法、病名、病名語尾、等
2) 各個数を文字数 1000 あたりの数に正規化
3) 各正規化個数とその全体合計により医学情報を判定
4) 論文表現種別個数を算出

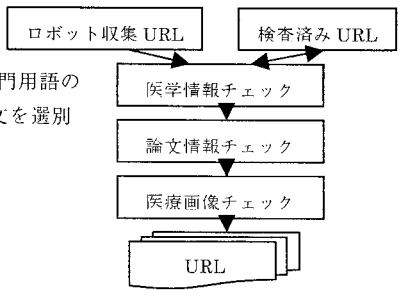

図2　医療情報フィルター

Filtering of Medical Data and Tag Application
Yuiko Suemura, Kiminori Simamoto, Tatsuya Iwashita　NS Solutions Corporation
Hiroshi Murayama, Harutoshi Ogai　Nippon Steel Corporation
Ritsuko Nakajima　Japan Science and Technology Corporation
Hirosato Nomura,Teigo Nakamura,Hidetoshi Nagai,Daigo Inoue,Daisuke Sakata　Kyusyu Institute of Technology

5) 種別の存在数により論文か判定

6) 画像を含むか判定（医療画像の分類は今後の課題）

4．医学論文抄録からのファクトデータ抽出

医療情報からのファクトデータ抽出の対象を医学論文抄録とした。日本医学放射線学会学術講演会抄録集の電子ファイルを研究用に提供いただき、抽出技術の研究を行った。情報抽出技術として、「テンプレートを用いた情報抽出」技術を中心に検討を行った。この方法は、文書に対する表層的な処理のみで情報が抽出できるため、大量の情報を高速に処理することが期待できる。実際には、更に幾つかの処理を付加して実現した。

4.1 抽出項目の設定

医学論文抄録は、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 1 に示す。その中から今回 21 項目の抽出を試みた。

抽出項目	タグ	説明
病名	BM	取り扱う主な病名
診断方法	SH	提案する診断方法
診断機器	SK	診断で使用する機器
対象症例	TK	診療対象の症例患者
症例例数	TR	症例、患者の例数
対象分類	STB	診療の評価対象
診療分類	SNB	診療方法の分類
評価項目	SNP	診療評価項目
評価数値	SNV	診療の数値評価値
評価全数	SNT	診療評価の全体例数
評価例数	SNS	診療評価の該当例数
評価補足	SNA	診断評価の補足

表 1　抽出項目の一部

4.2 医学専門用語辞書による医学論文抄録タグ付け

「テンプレートを用いた情報抽出」技術では、事前に抽出項目をタグ付けした「テンプレート学習用文書」を用い、情報抽出処理に使用する「テンプレート」を作成する。「テンプレート学習用文書には、抽出すべき情報を特定、識別するためのタグを付加する。タグ付けのフローを図 3 に示す。その中で、形態素解析ツール JUMAN に医療用の専門辞書を作成登録して、タグ付けしている。また、この専門辞書作成を支援する機能として、医療用語抽出分類システムを開発し、入力文中から専門用語の分類対象となる語を抜き出し、パターンマッチ、スコアリングの手法を用いて入力文中の未知語、複合語の分類属性を決定し、辞書に登録している。なお、タグ付け文書はＸＭＬの形式に準拠させている。

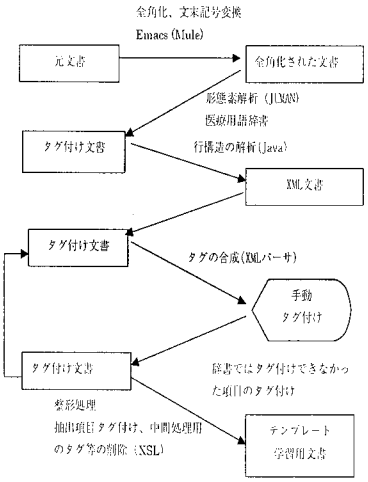

図 3　タグ付けフロー

5．医療情報収集システムの評価

医療情報収集システムの評価として、医療情報フィルタの選別性能、タグ付け精度について報告する。

6．まとめ

「高度医療ネットワークに関する調査研究」の一部として実施した医療情報収集技術の成果の中で医療情報のフィルタと医療抽出項目の自動タグ付けについて報告した。残された課題は、1) タグ付け抽出項目の拡大とタグ付け精度の向上、2) 大量医療情報の収集、3) 医療画像の分類技術開発等である。これらの課題について、さらに研究を継続していく予定である。

最後に、「高度医療ネットワークに関する調査研究」は、平成 10 年度から科学技術振興調整費に基づいて行われてきた。医学論文抄録をご提供いただいた日本医学放射線学会、並びに高度医療ネットワークの研究にご協力いただいた関係各所の皆様に深く感謝させていただく次第である。

参考文献

[1] 岡野弘行、大形英男、大貝晴俊、小長谷幸：情報収集技術の研究、平成 10～12 年度科学技術振興調整費調査研究報告書

[2] 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 2001-NL-141,pp103-110,2001

[3] 大貝、大形、中島、嶋本、寿恵村他：医療情報収集とファクトデータの抽出、情報メディア学会研究会、2001