（別紙１）

# 論文要旨

# 論文題目　画像診断報告書からの所見抽出

学際情報学コース
27102　今井　健

指導教官　小野木雄三

**研究背景と目的**

近年、病院情報システム内において日々電子的に蓄積されていく診療情報が飛躍的に増加している。しかし、これらの多くは自然言語（日本語）にて記載されており、データマイニングや知的検索、診断支援、あるいは教育、経営など様々な目的に２次利用するために、適切な構造化手法の確立が求められている。

情報を入力する際に予め定められたコードを用いるという方法も考えられるが、診療現場において入力された情報をコーディングする専門家 (Transcriber あるいは Librarian) が存在する欧米と違い、我が国では医師が自ら情報を入力しているため、負担を考えると現実的ではない。さらに、最近ではキーボードからの入力だけでなく、連続音声認識技術により、マイクからの音声入力が直接文章に変換されて記録されるようになってきており、自然言語で蓄積される診療情報はますます増加する一方である。そのため、診療情報の構造化には計算機による何らかの支援が必要である。

欧米の医学・医療分野では用語体系の整備が進んでおり、これらを用いた自然言語処理の研究が盛んに行われている。しかし、我が国においては日本語医学用語における概念や知識の整理が未だ十分に実現されておらず、日本語医学用語など自然言語処理の基盤となるリソースがようやく近年徐々に整備されてきた段階である。

そのため我が国の医学・医療分野においては従来主に用語解析の研究が成されてきた。また文章からの意味抽出の観点からはＳＧＭＬ化されたテキスト・医学教科書を対象にしたものなどが存在する。しかし、実際の診療情報テキストでは「記入医による表記の多様性」や「臨時一語の出現」などの特徴のため、まず複数の「正規化あるいは標準化」を行う前処理を適用しなければならず、そのような実際の医用文書を対象とした自然言語処理研究はほとんど行われていない。

本研究ではこのような背景のもと、診断支援システム等への２次利用のため、実際に用いられて

いる診療情報テキストの構造化を目指したものである。そのための第一段階として、そのようなテキストの代表例である画像診断報告書を対象とし、画像診断において重要な「所見」に関する記述を抽出する手法を構築し、その有効性の評価を行った。ここにおいて「所見」とは「ある部位における病変や異常（以下「所見要素」）の存在の有無、あるいは状態に関する記述」のことで、もし存在すれば属性情報も伴う。

**実験材料と方法**

本研究では、東京大学付属病院にて2003年2月までに電子的に蓄積されたCT及びMR画像診断報告書のうち、最近の20,000件のものを対象とした。これを10,000件ずつ排反に分け、1つのセットはルール構築用、もう片方のセットをテスト用とした。

意味解析や知識の抽出などに自然言語処理分野の知見を援用するためには、そのルールの元となる頻度情報が必要であるが、それには正解タグが付与されたコーパスが必要である。しかし診療情報でそのようなコーパスは未だ存在せず、上記リソースも生コーパスである。

一般的な自然言語処理分野における解析手法としては、タグ付きコーパスなどからの学習により、頻度情報に基づく構文解析や格フレームの利用などの研究がされているが、本研究対象ではそのようなリソースが存在しない。また新聞などのリソースから構築されたルールと違い分野特有の表現が多く出現することの理由により、本対象分野に適した手法の構築が必要である。そのため、一般に最も精度が高いと言われている用手的な手法によりシステムを構築した。その際、画像診断報告書の特性や所見の抽出という目的を考えた時、優先するべきと考えられる要請は以下の通りである。

(1)「右下腿内側多嚢胞性腫瘤」など複数の語を連接させた臨時一語などを含め、医学用語を正しく認識すること。(2) 画像診断領域で一般に重要と言われている「所見に関して確定的に記述されている単文的な文章」をきちんと解析できること。(3) 医学・医療分野では重要な「数値・記号表現」を正しく認識すること。

これらの点を踏まえ、以下の4つの手順でシステム構築を行った。

Step1) 正規化処理と形態素解析

形態素解析を行う前に、生コーパス中の文章を1文毎に分割しなければならないが、文章の区切り記号の多様性や文中の数値・記号表現との区別が困難である問題が存在する。そのため、数値・記号表現のタグ付け処理、各文毎への分割処理を行い、その後で形態素解析処理を行った。形態素解析にはJUMANを用い、既存の統制用語集である医学用語シソーラス第5版、ICD10対応標準病名マスター、からの71,253語に加え、手動で構成した画像診断領域特有の辞書3,413語を形態素解析用の辞書に追加した。

Step2)IDIOM化処理

本論文では、IDIOMとは複数の形態素を統合した、より大きな意味上のまとまりのこととする。この段階では、Step1の結果複数に分割されてしまう臨時一語などの医学用語を結合し、【部位名】や【所見要素】などの意味属性を付与した。また、文末の「存在の有無」や「状態の記述」を表す

【主張句】を1語に統合した。さらに、意味上の要素をなるべく大きく抽出するために、数値・記号表現の属性化、並列関係処理、近接係り受け処理など、再帰的な統合処理を行った。

Step3) カテゴリフィルタ

画像診断報告書中の文章はその内容によって「所見」「検査内容」「比較対象」「病状」「リコメンデーション」などのカテゴリに分類される。文末における統語的なフィルタリング処理によって、「所見カテゴリ」の文章だけを選別し、所見抽出の対象とした。

Step4) 文型パターンを用いた所見の抽出

この段階では日本語を含み、また単文的な構造を持つ所見文を対象とし、所見抽出を行った。まず、各文章に対し「部位」「所見要素」の後に取り得る助詞と「文末の主張句」の組み合わせに基づくメタ情報である「格に基づいた文型パターン」を考えた。さらに、これが所見記述を表す際の最も簡単な構造である「自明な単文」という概念を導入し、文型パターンとのマッチングにより、最終的な所見抽出を行った。

**システムの評価**

本システムの各処理の性能を、医学知識を持つ専門家によってつけられた正解との比較実験により評価した。尚、評価実験の際は、肝臓に関する所見を含むテスト用セット3,370件のものから、ランダムに抽出した30件中の329文を対象とした。

Step1からStep3のまでの各処理についての成績を以下のTable1に示す。

| | 処理内容 | Recall | Precision |
|---|---|---|---|
| Step1 | 数値・記号表現のタグ付け | 99.5% | 99.0% |
| | 文章の分割 | 100% | 100% |
| Step2 | 文末主張句の抽出 | 84.5% | 100% |
| | 部位名・所見要素の抽出 | 86.9% | 98.4% |
| Step3 | カテゴリフィルタ処理 | 100% | 97.2% |

Table 1　Step1から3までの主な評価実験結果

上記のように、いずれも高精度の処理が可能であることが示された。またこれらはルール構築用セットにおける成績ともほぼ同様であり、矛盾しないという結果を得た。

次に、Step4の「格を用いた文型パターンによる、所見抽出」であるが、Step3までの処理を経て抽出された、日本語を含む所見文章258文の文型パターンに対し「自明な単文」は24種類構成された。これを用いたパターンマッチングによる所見抽出の成績はRecall=44.6%, Precision=100%であり、これはルールセットに対する成績（Recall=45.2%, Precision=100%）と矛盾しない。

また、今回は所見に関して単文的な構造を持つ所見文を対象としている。このようなものは207文存在し、対象を「単文所見」に限定したRecallは55.6%であった。

### 考察とまとめ

テストセット中の全329文中、「日本語を含む所見文」は76%、さらに「単文的な所見」は63%である。また、全150個の文型パターン中、自明な単文構造を用いて生成されたわずか24個(16%)の文型パターンのみで、単文的な所見に対し精度100%のまま55.6%の再現率を実現することができた。

これにはStep1から3までの処理が文型の正規化あるいは簡単化に大きく寄与したためであると考えられるが、これらの主要な処理の効用を「その処理を除外した時にどの程度Recallが低下するか」という基準で調べた。その影響度をTable2に示す。

| 除外した処理 | Recall | Recallの低下 | 全文型パターン数 | 所見抽出ルール数 |
|---|---|---|---|---|
| 【並列・係り受け統合】 | 23.7 % | - 31.9 p | 172 (+ 22) | 16 (- 8) |
| 【数値・記号表現のタグ付け】 | 24.6 % | - 31.0 p | 160 (+ 10) | 19 (- 5) |
| 【属性化(数値・記号・形容詞】 | 33.3 % | - 22.3 p | 159 (+ 9) | 19 (- 5) |
| 【複合語候補の特定】 | 46.4 % | - 9.2 p | 155 (+ 5) | 19 (- 5) |
| 本システム(参考) | 55.6 % | | 150 | 24 |

Table 2　所見抽出に対する各処理の影響

これらは所見抽出の前段階として構築したStep1から3中の主要な処理であるが、いずれも最終的な所見抽出に大きな影響を及ぼす重要な処理であることが定量的に示された。

また今回作成した画像診断領域の辞書(3,413語)についても同様に除外して実験したところ、Recallが37.7pも低下した。これは、その他追加した統制用語集71,253語よりも量としてはかなり少ないが、正規化処理やIDIOM化処理よりも重大な影響を持つことが判明した。

さらに、再現率の向上に向けては、(1)文頭における典型的な表現 (2)英語辞書の追加 (3)文末主張句の3つの簡単な改良を施すだけで、特にアルゴリズムを変更することなく、7割程度までRecallが改善されることが示唆された。しかし、これ以外のものについては、「部位IDIOMの高度化」や「倒置表現」「複雑な係受け構造」など、本システムのアルゴリズムや手法自体をより高度に改良する必要があることが判明した。

本システムでは、所見に関する単文構造に関してのみ抽出を行った。単文構造の抽出精度を高めるためには、上記の考察が参考にできると思われ、今後の課題である。一方で次の段階として、複文・長文の類を対象にする必要があるが、単文に対する文型パターンの知見が十分に得られた後、文中の主張句を考えることで、単文に対する文型パターンを再帰的に適用すれば、単純な複文・重文構造を解析することが可能になると思われる。

また、抽出された所見情報における「部位などの情報欠損」「ゼロ代名詞問題」については、何らかの照応解析が必要であるが、これについては箇条書き段落における部位のスコープを適用できる可能性がある。

また本システムの対象を一般化し、他の医用文書を対象とする観点からは、医学・医療分野においてはほぼ共通に使われ、かつその中でのジャンルに依存しにくい「数値・記号表現」についての

「正規化処理」が適用可能である可能性が高く、今後の検証が必要である。また、インシデントレポートや、病理検査レポートなどでは、画像診断領域とは異なる分野特有の表現が使われるため、本研究における分野固有辞書 (3,413 語) のような辞書を統制用語集に追加する必要があるだろう。これは本研究における分野固有辞書の効用の大きさから見ても重要であると思われる。

本研究で取り上げた「所見」は医療分野における多くの診療情報テキストに共通する重要情報であり、所見を構成する要素に大きな差異はないため、電子カルテや病理検査など他の報告書からの所見抽出にも同様の手法が適用できると考えられる。その一方で、インシデントレポートのような対象は、所見を記述する「部位」＋「所見要素」＋「主張句」など、分野に特有な頻出の言い回しが少なく、タグ付きコーパスを用いた機械学習などの手段を考えなければならないだろう。しかし、本研究のような医学用語認識、属性表現抽出の手法を適用することで、コーパス作成の支援に寄与することができ、その後では一般的な自然言語処理の知見が援用できる可能性がある。

従来我が国の医学・医療分野において、画像診断報告書など実際の診療情報テキストを対象とした情報抽出はほとんど行われて来なかった経緯があるが、必要なリソースを含め１から処理体系を構築した本研究の成果は今後ますます盛んになると思われる医学・医療分野での自然言語処理研究の突破口となる重要な基盤を築くものである。