（別紙２）

論文審査の結果の要旨

論文提出者氏名　今井　健

近年、診療文書は紙カルテに代わり、電子的に蓄積されるようになってきている。今後は蓄積された情報を利用して診療や研究に応用することが期待されているところであるが、診療文書の多くは自然言語で記述されているために、情報を抽出することは困難とされている。本論文は、日本語の自然言語で記述された診療文書の中でも比較的重要度の高い画像診断報告書から所見を抽出する手法を、東大病院で蓄積された 2 万件の報告書を材料として自然言語処理の技術と用手的ルール構築により開発し、その成績を評価したものであり、下記の結果を得ている。

1. 辞書の構築と意味属性の付与

日本語医学統制用語集（日本医学用語シソーラス第 5 版、ICD10 対応電子カルテ用標準病名マスター）71253 語に対し「医部位」「疾患名」「検査手技・機器名」3 種の意味属性を付与し、蓄積された画像診断報告書から抽出した語彙 3413 語に対し「医変化」「医状態」「医所見」「医修飾」など、計 29 種の意味属性を付与した。併せて 74666 語の辞書を構成した。

2. 正規化処理

画像診断報告書の文体・記法は医師や検査種別によって異なるため、文章を切り出すためには、表記の揺れを統一し日付・数値・記号・省略・箇条書きなどをタグ付けする必要がある。ルールセット 1 万件から用手的に計 191 個のルールを構築した。テストセットから肝臓に関する記述を含むレポート 3370 件からランダムに抽出した 30 件(329 文、数値・記号表現 195 件)に対する成績は、数値・記号表現のタグ付けでは精度 / 再現率(以下 P / R と記載)は 99.0% / 99.5%、文章切り出しでは P/R ともに 100%であった。

3. IDIOM 化

画像診断報告書の所見記述内容は【部位】【所見要素】【主張】という意味内容に分類される。例えば「肝左葉に大きさ 3cm の腫瘍が存在する」の場合、[肝左葉【部位】] に [大きさ 3cm の腫瘍【所見要素】] が [存在する【主張】] となる。それぞれの意味内容は複数の形態素の連接であるため、これらを併せて意味上のまとまりにする処理を IDIOM 化処理と呼ぶ。これは以下の４つ(A-D)の処理に分けられる。

A. 医学用語の IDIOM 化

① 複合語候補の生成：

医学用語の多くは漢字とカタカナの連接であることから以下のルールを構築した。

・漢字のみ・かつ動詞でない場合に結合する

・1 文字以上のカタカナと漢字で構成され、かつ名詞である限り結合する

上記テストセット(複合語候補 128 件)での成績は、99.1% / 89.8%であった。

② 複合語候補の属性決定処理：取り出した複合語が【部位】【所見要素】のいずれの意味内容に分類されるかを、以下のルールで決定する。

・部位関連属性を持った形態素が最後の場合は【部位】

・所見要素関連属性を持った形態素が最後の場合は【所見要素】

・「医部位」の後ろが「医修飾」「一般名詞」「接尾辞」「形容詞」のみなら【部位】

・「医変化」「医状態」「医所見」の後ろが一般名詞なら【所見要素】

・「医疾患」の後ろが「サ変名詞」か「医修飾」なら【所見要素】

テストセットでの成績は 98.0% / 85.2% であった。

B. 文末主張句の IDIOM 化

① 文末主張句の表面的な肯定・否定表現の特定

文末の動詞句・形容詞句に関して「認める」「疑われる」「正常だ」など 26 種のルールを頻度順に構築、さらに文末の疾患名言い切りに対するルールを追加して表面上の肯定・否定を識別した。また表現に応じて 3 段階の確信度を付与した。テストセット 329 文に対し、肯定・否定タグが正しく付されたものは 66%、言い切り型を含めると 71.3%、さらに本研究の目的である日本語を含む所見文を対象とすると 94%が回収された。成績は 100% / 94.0%。

② 文末主張句の静的状態・変化・存在への分類

以下のルールで文末を処理する。

・変化：「医変化」「医所見」属性の【所見要素】＋<肯定・否定>

・状態：「医状態」属性の【所見要素】＋<肯定・否定(言い切りを含む)>

・存在：上記以外の<肯定・否定>

テストセットで日本語を含む所見文に対する成績は 100% / 84.5%であった。

C. 数値・記号表現・形容詞の属性化

【部位】【所見要素】の属性として数値・記号表現・形容詞を包含させることにより、IDIOM をより大きなまとまりにする処理である。例えば「左肺上葉 S1+2」では「左肺上葉【部位】」に「S1+2」という肺区画を示す記号が【部位】の属性として取り込まれる。「医修飾」に分類される修飾語 336 語を用いて、以下のルールを構築した。

・記号表現：【部位】＋「記号表現」　→　【部位】に統合

・数値表現：「数値表現」＋「の」＋【所見要素】　→　【所見要素】に統合

・「医修飾」＋「に、の、な」＋【所見要素】　→　【所見要素】に統合

・「医修飾、形容詞、連体詞」＋【所見要素】　→　【所見要素】に統合

テストセットに対する成績は、記号・数値は 100% / 93.9%、形容詞は 96.3% / 91.9%。

D. 並列IDIOMの再結合、部位同士・所見要素同士の係り受けの統合

① 並列関係：【部位】同士、【所見要素】同士が、(「や」「と」「、」「・」「及び」「および」「ないし」) で結ばれている限り結合し、ひとつのIDIOMとして統合する。

② 近隣係り受け：例えば「局所再発を疑わせる腫瘤性病変」の場合、2つの【所見要素】のうち後者が主体で前者は修飾であり、これを1つの【所見要素】IDIOMとして統合する。ルールセットでは2つの【所見要素】の間に挟まれる形態素の数は3個以下であったため、これらのパターンを頻度順に収集してルールを構築した。また【部位】における係り受け関係は間に1形態素を含むものだけであり、頻度最上位の「の」で結ばれるものに対するルールを作成した。テストセットに対する成績は100% / 73.1%。

IDIOM化の総合成績：Bについては単独に評価可能で100% / 84.5%、ACDには依存関係があるため累積的に評価して98.4% / 86.9%であった。

4. カテゴリフィルタ

画像診断報告書の全体は「検査内容」「比較対象」「病状」「所見」「リコメンデーション」のカテゴリに分けられる。所見抽出を行うためには所見カテゴリを正しく識別する必要があり、これをカテゴリフィルタで行う。文末の典型的な言い回しを利用して34個のフィルタを構築し、テストセットに対する成績は100% / 97.2%。

5. 格を利用した文型パターンに基づく所見抽出

以上の処理により、所見を記述した文章はIDIOMと助詞だけで構成されたIDIOM列に変換されている。画像診断報告書で所見を記述する文章の多くは単文的な文章であることに着目し、3種類のIDIOMがそれぞれ、高々ひとつしか含まれないIDIOM列パターン(自明な単文)を識別するルールを構築した。ルールセットでの再現率は45.2%、テストセットでは44.6%、精度はいずれも100%であった。またテストセットで単文のみを対象とした場合の再現率は55.6%であった。

6. 各処理の効用

本システムを使った所見の抽出成績は55.6%であるが、各処理を除いた成績によって個々の効用を調べた。並列・係り受けでは23.7%、数値記号表現のタグ付けでは24.6%、数値記号・形容詞の属性化では33.3%、複合語候補への属性付与では46.4%に成績が低下した。また本研究で構築した辞書を除いた場合は17.9%であった。辞書が最も大きな寄与を、次に並列・係り受け処理および数値記号表現のタグ付けが重要であることが示された。

以上に基づき、文頭のIDIOM化失敗例への対策、IDIOM化に影響する語彙の辞書への追加、文末主張句ルールの改善に関する簡単な改良を行った場合の成績を推定することにより、再現率を71%にまで改善できることが示された。

以上、本論文では自然言語で記述される画像診断報告書からの所見内容抽出を 100%に近い精度を維持しながら再現率を向上する方針で用手的にルール構築を行い、最終的に 55.6%の所見を抽出するという結果を得ている。これは日本語自然言語で記述された画像診断報告書からの情報抽出としては初めての成果である。また新聞や出版物などのきれいな文章ではなく、複数の医師が様々なスタイルで記述している変異の多い文章を対象としているにもかかわらず、一定の成績を出していること、特に診療においてとりわけ重要な記号・数値表現を高率に抽出している点は高く評価できる。また本研究の手法自体は Semantic Grammar に分類され、現在の自然言語処理における最新の技術ではないが、日本の医療分野における自然言語処理の土台と位置づけられる。特に今後、医療分野で自然言語処理システムを開発していくためには、機械学習の材料としてタグ付きコーパスの構築が必須であるが、本研究の成果により生コーパスからタグ付きコーパスへの変換を支援するツールを作ることが可能である。その点でも本研究は今後の研究の土台となるものである。

本論文は、医療情報学(医学)と自然言語処理(理工学)との境界領域の研究である。自然言語処理の立場からは、言語処理の要素技術を現実の医療分野のテキストに適用することに関して、博士の価値がある新規性のある研究である。医療情報学においては、電子カルテをはじめとして医療の電子化が進み、医学専門用語を含んだ日本語テキストから情報を抽出する技術が重要になりつつある状況下で、日本の医療情報学に大きなインパクトを与える新規性のある研究である。

以上により、本審査委員会は、本論文が博士（学際情報学）の学位に相当するものと判断する。