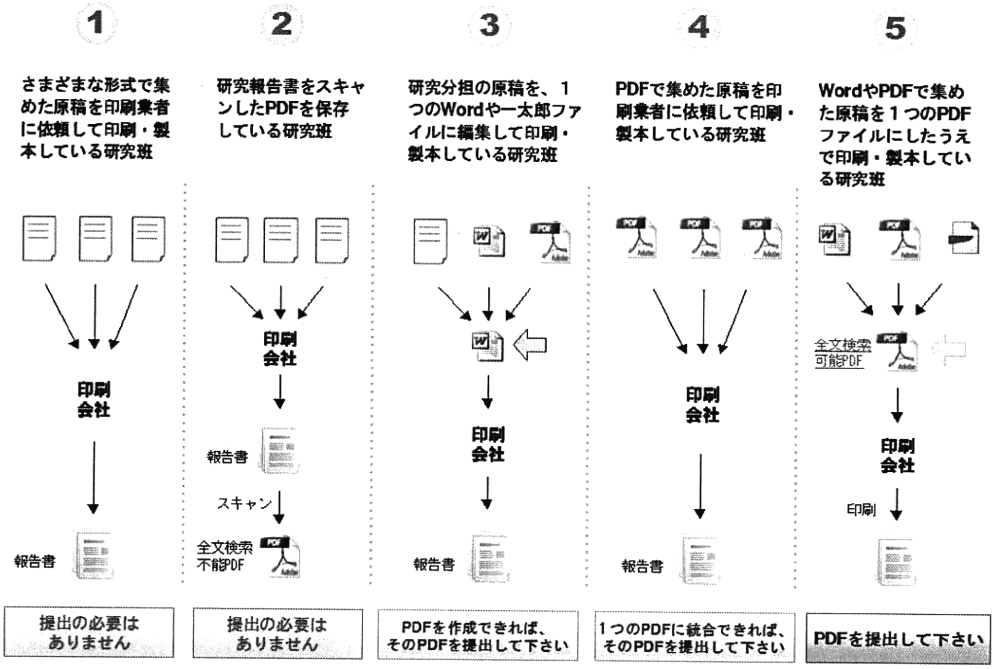


図 1 回収対象ファイルの図示

本研究分担においては、過去の研究報告書の効率的なインデックス化に向け、研究代表者に依頼することにより紙媒体の報告書の元となった電子データの回収を試みた。また、今後の円滑なファイル回収に向け、ファイルアップローダやメール一斉送信システムを用意した。しかしながら、回収率が 1.4%と極めて低いうえ、実際に利用が可能なデータはさらにその 4/10 と、実際の事業に用いるには多くの困難があることが明らかとなった。

まず、過去の研究課題の場合、研究代表者への連絡自体に困難が生じる。さらに、研究代表者に連絡がついたとしても、研究代表者にとっても相当な過去の課題の場合、原稿やデータが散逸してしまっている可能性が高く、回収率の低さとなって現れることになった。一方で、これらの問題に関しては、最近の研究課題に近づくに連れて次第に改善されていくことが期待される。提出ファイル数、提出ページ数は、年度を経るにしたがって増加して来ているため、たとえ 10%の有効回収率だとしても、OCR 化に必要な費用を大きく節約できる可能性がある。今年度の研究では時間的な制約が強く複数年度のファイル回収が困難であったが、今後、対象年度を変えて回収を続けることにより、回収率の増加を確認していく必要があるだろう。

また、今回、ファイルの回収に成功した 10 研究課題のうち、実際の利用に適うデータが 4/10 と少なかった。不適とされた報告のうち、2 例は総括研究報告書のみの不完全な提出であり、2 例は全文テキストが含まれないスキャン PDF の提出であった。また、厚生労働科学研究の成果報告書として求める書式を満たさない報告書も 2 例あった。こうした事態の背景としては、今回の実験に利用した説明文章が冗長で、どういうファイルが回収対象であるかの伝達が行えていなかったことが指摘できる。今後は、たとえば図 1 のように、どういうファ

イルを回収対象としているのかを図示するなどの工夫が欠かせないだろう。
また、研究成果データベースの全文テキスト化を行っていくうえでは、過去の研究成果報告書の全文テキスト化だけでなく、今後、提出して頂く報告書の全文テキスト化も同時に考慮する必要がある。その意味では、今回いくつかの研究班が複数の PDF ファイルを提出したことで明らかとなったように、それぞれの研究班は各自の事情に応じた報告書の作成スタイルを有しているため、ファイル回収事業は、現状どのような形式で研究報告書を作成しているのかを明らかにする好機であるとも言える。今後は、回収対象ファイルを分かりやすく示すことに加えて、報告書の作成スタイルに関する調査を兼ねることにより、PDF 化に適した作成方法を行っている研究班の数やどのような形で報告書を作成しているかの傾向を把握することが望ましい。これにより、ファイル回収策自体を改善していくことが可能となるだけでなく、今後、研究成果報告書を電子ファイルにより提出させるための基礎資料として用いることが出来るだろう。また、回収ファイルの精度向上を客観的に把握していくうえでは、回収したファイルが期待しているファイルであるかを検証する必要がある。そのためには、紙媒体で提出された研究成果報告書の PDF とアップローダにより集めた PDF を効率的に比較するためのツールの検討なども求められるだろう。

D-2) 研究報告書の OCR 化に関する動向

一般的に、OCR 処理における認識精度は、読み取り対象となる画像の状態に大きく影響を受ける。たとえば、Fax のように最初より 2 値化された画像では、OCR 性能の向上を目指した画像処理を行うことが困難である。一方、スキャン画像が 400DPI 相当以上のグレースケール画像かカラー画像であれば、ノイズ除去や、OCR 処理のための 2 値化処理などの画像処理手法を容易に用いることが可能である。現在の GRANTS システムでは、2005 年に至るまでは 200DPI でのスキャンを行っており、2005 年度以降のスキャンも 300DPI に留まっている。国会図書館なども、現在では、OCR 処理を前提とした 400DPI のカラー画像での書籍スキャンを行っているようであり、参考となるだろう。

OCR 化テキストの精度向上に関しては、現在、書籍のスキャンから OCR を利用したレイアウト解析、書籍の構造解析、テキスト化、テキストの読み順の自動指定、電子書籍フォーマットへの変換を行い、Web ベースで分散して確認修正を行うようなシステムが利用されている。ヨーロッパでは、圏内に存在する膨大な量の貴重な古文書をスキャンしてデジタル化するために、IBM と欧州連合(EU)が協力し、欧州各地の図書館や研究所、大学などが保有する書籍や文書のデジタル化も進んでいる[1]。今後、こうした試みの日本語版を発展させることにより、既存資料の全文テキスト化をより効率化していく必要があるだろう。

D-3) 研究成果報告書の全文テキスト化に要するコスト

GRANTS の検索機能高度化を実現するためには、過去の研究成果報告書を全文テキストデータ化するとともに、今後、提出される研究成果報告書もまた全文テキストデータ化していかなければならない。これらには、当然、予算が必要となる。
今年度の研究において、1998 年から 2002 年度分の 374,113 ページを処理するために、

[1] http://www.impact-project.eu/home/

2,500 万円が必要であった。これは、単純計算でページ単価 66.8 円となり、約 80〜100 円という OCR 処理の市価より安価である。その理由として、研究報告書には読み取りに適さない図表や見出しページのように処理をほとんど要さないページが多く含まれていた点が考えられる。

次に、残されている 2003 年から 2009 年度の 1,991,675 ページを処理する費用を見積もる。現在の費用モデルを単純に適用すると、下に示す通り、13,000 万円ほどの費用が必要となる。次に、ファイル回収事業を継続し有効回収率を平均で 10%まで改善し、また、技術進歩と大量発注により OCR 処理の単価を 50 円まで低廉化できるとすると、およそ 9,000 万円となる。いずれにせよ、既存の報告書のデータ化だけでおよそ 1 億円を要するため、今後、毎年 1,000 万円ずつ電子化予算を掛けたとしても、全データの処理に 10 年が掛かることになる。

現在モデル

1,991,675 ページ × 0.99 (OCR 率) ×
66 円 (円/ページ) = 130,136,044 円

技術進歩モデル

1,991,675 ページ × 0.90 (OCR 率) ×
50 円 (円/ページ) = 89,625,375 円

これらは、収録済みファイルの電子化費用であり、この他に毎年提出される研究成果報告書の電子化についても合わせて検討する必要がある。まず、2010 年度以降の研究成果報告書を電子ファイルにより提出させるとすると、これは事実上、研究報告の「電子出版化」に他ならない。100%の回収率とするためには、提出を義務化する必要があり、そのためには各研究課題に支払う補助金に電子出版化の予算を上乗せしなければならないだろう。研究報告書は、年間 1,600 課題ほどあるため、各報告に印刷費 10 万円を積むと、それだけで 16,000 万円の追加予算が必要となる。

一方で、年間増加分の 30 万ページ分を OCR 化するためには、今年度の予算モデルでページ単価を 66 円とすると、下記のように 2,000 万前後と大幅に予算を削減することが出来る。これは、各報告書で見ると 12,000 円前後となり、電子出版化を義務化した際の費用よりも大幅に安価であることが分かる。

提出義務化

1,600 課題×100,000 円= 160,000,000 円

提出後処理

300,000 ページ×66 (円/ページ) =
19,800,000 円

さらに、以上の処理に加えて、PDF ファイルの回収による効率化分を加味してみよう。過去ファイルでは、平均の有効回収率を 10%と仮定したが、直近の研究課題に関しては回収率の向上が見込めるため、10%から 50%の範囲で改善できると仮定しよう。以下のように、回収率向上によって直線的に予算の削減が可能となり、50%の回収率により、およそ 1,000 万円での電子化が可能となる。電子提出にインセンティブを設けることで回収率を向上させ、100%の回収が可能となれば、当然、電子化予算は不要となる。

回収率 10%

300,000 ページ×0.90(OCR 率)×
66 (円/ページ) = 17,820,000 円

回収率 50%

300,000 ページ×0.50(OCR 率)×
66 (円/ページ) = 9,900,000 円

以上より、今後の電子化費用は、全文テキスト付き PDF 報告の回収率向上により極小化できることが分かる。また、そのためには、過去の研究報告書の電子ファイル回収を継続し、手法を改善していくことにより回収率を向上させていくことが重要である。ただし、このコストモデルは、PDF ファイルの回収率だけでなく OCR 読み取り単価にも依存しており、読み取り単価を大きく削減する技術革新により更なる低価格化が行える可能性がある。

E. 結論

厚生労働科学研究成果データベースの検索機能強化に向けて、全文データの効率的な確保手段が求められていた。そのためには、紙媒体で提出された報告書を OCR 処理により読み取るだけでなく、提出された報告書の元となった電子ファイルそのものを研究代表者側より入手する方法がある。そこで、本研究分担では、過去の研究報告書を効率的に全文テキスト化し、インデックス化するための基礎資料とするため、研究代表者への依頼に基づくファイル回収実験を行い、今後求められるコストの推定を試みた。

今年度においては、研究開始時期の問題があり、1998 年度の研究代表者のみを対象としたファイル回収を試みた。その結果、回収率は全体の 1.4%に留まり、実際の利用に足るファイルの回収率は 0.57%に過ぎなかった。回収されたファイルには意図したものと異なるファイルも多く、ファイルの回収率を上げていくためには、アップロード対象としているファイルを簡潔に伝える図の提示などの工夫が望まれた。

厚生労働科学研究研究成果データベースの全文テキスト化に際しては、今後、過去ファイルの回収だけでなく、最新の研究成果報告書の電子ファイルによる提出も必要となっていく。そのためには、そもそも研究代表者がどのような形態で研究報告書を作成し、提出しているかという基本的な情報が欠かせない。そうした情報により、より実態に即した形式での電子化ガイドラインの作成が可能となり、また、効果的なファイル回収が可能となるだろう。そのためには、今後のファイル回収において、研究成果報告書の作成スタイルに関する調査を行いつつ、回収対象となりうる作成方式を取っている研究班を対象とした選択的なファイル回収を行うなどの工夫が望ましい。

今回の試算では、未だ全文テキスト化が行われていない 2003 年から 2009 年度の処理のために、9,000～13,000 万円の費用が必要となることが明らかとなった。今後、限られた予算で研究成果報告書の全文テキスト化を行っていくためには、ファイル回収策自体を改善しながら回収事業を進めていくことにより、ファイルの回収率を上げ、OCR 処理が必要なファイル数を削減していく必要がある。また、OCR 処理そのものを効率化することにより、さらに低廉化しうる可能性がある。たとえば、OCR の機械処理の精度を上げることがコストダウンに繋がるために、現状よりもスキャンの解像度を上げ 400DPI にすること等が求められる。

F. 研究発表

なし

G. 知的財産権の出願・登録状況

なし

厚生労働科学研究費補助金(特別研究事業)
分担研究報告書

OCR 処理された文献に関する知的検索研究

研究分担者　岡崎 直観(東京大学 大学院 情報学環)

研究要旨

厚生労働科学研究では、毎年約 1,500 課題の研究を実施しており、その研究成果報告書が年々蓄積されている。従来の GRANTS システムでは、研究報告書の概要や研究者名などのメタデータのみが検索対象となっていたが、本プロジェクトで開発する iGRANTS では、OCR で電子化された報告書の本文を検索できるため、インデックスされている情報の量が飛躍的に増加した。一方、検索対象の情報量が増えたことで、検索クエリに対して数百～千件の成果報告書が見つかることも珍しくなくなった。

そこで、検索クエリの入力や検索結果の把握を支援するため、本研究では、検索システムにキーワード、類似キーワード、類似報告書の視覚化及び検索など、知的検索機能を付与する手法を検討した。具体的には、OCR 化された研究報告書に対し、テキストマイニング技術を適用し、キーワード、類似キーワード、類似報告書の情報を自動的に獲得する手法を開発した。自動獲得された情報は、全文検索システムのメタデータとして登録され、iGRANTS システムに統合される。iGRANTS の検索結果の画面では各報告書のキーワードが表示され、ユーザーは報告書の中身を即座に予測することが可能になった。また、類似キーワードや類似報告書を視覚化する機能により、ユーザーが探したい報告書へ積極的にナビゲートすることが可能となった。本研究で開発されたシステムを運用することにより、知的検索機能に対する要望を調査し、成果報告書データベースがユーザーに積極的に提示すべき情報や、そのインタフェースに関して、検討が進められると期待される。

A．研究目的

厚生労働科学研究では、毎年約 1,500 課題の研究を実施しており、その研究成果報告書が年々蓄積されている。この膨大な研究成果報告書に対し、全文検索を実装することで、ユーザーの検索クエリに適合する報告書を検索できる。従来の GRANTS システムでは、研究報告書の概要や研究者名などのメタデータのみが検索対象となっていたが、本プロジェクトで開発する iGRANTS では、OCR で電子化された報告書の本文を検索できるため、インデックスされている情報の量が飛躍的に増加した。

一方、検索対象の情報量が増えたことで、検索クエリに対して数百～千件の成果報告書が見つかることも珍しくなくなった。大量の検索結果が見つかった場合、ユーザーが自分の検索要求にマッチする報告書を一つ一つ選別しなければならない。検索結果を見ながら、検索クエリを改善し、検索結

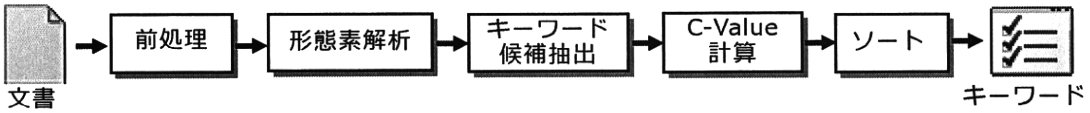


図 1 キーワード抽出の流れ

果を絞り込むことも必要であろう。さらに、ユーザーが適切な検索クエリを入力できない場合も多く、検索クエリが間違っている、検索クエリと報告書中の表現と一致していない、検索クエリが曖昧である、等の問題にも直面する。

本報告書では、ここに挙げた課題を解決し、検索クエリの入力や検索結果の把握を支援するため、検索システムにキーワード、類似キーワード、類似報告書の視覚化及び検索機能を付与する手法について述べる。キーワードとは、研究報告書を特徴づける専門用語のことで、検索結果の各報告書を短い表現で表すキーワードを提示したり、キーワードに基づく検索を実現するものである。類似キーワードとは、あるキーワードと関連の高い別のキーワードを提示し、「もしかして？」のようなキーワード間のナビゲーションを行う機能である。類似報告書とは、検索された報告書と似ている報告書をサジェストする機能である。これらの機能を、本研究では知的検索機能と呼ぶ。

これらの知的検索機能を実現するには、各報告書からキーワードを抽出し、類似キーワードや類似報告書の認定を行わなければならない。本プロジェクトにおいてインデックスされている報告書の量は膨大であるため、今後新たにシステムに追加される報告書に対応するには、上記の処理を自動化する必要がある。そこで本研究では、OCR 化された研究報告書に対し、テキストマイニング技術を適用し、キーワード、類似キーワード、類似報告書の情報を自動的に獲得する。自動獲得された情報は、全文検索システムのメタデータとして登録され、iGRANTS システムに統合される。

B．研究方法

B-1）キーワード抽出

本節では、OCR 化された研究報告書に含まれるキーワード(専門用語)を、自動的に抽出するアルゴリズムについて述べる。アルゴリズムの概要を図 1 に示す。本アルゴリズムは、報告書毎に実行され、各報告書を特徴づける研究キーワードを抽出する。以下、図 1 に示すアルゴリズムの構成要素毎に、処理内容を説明する。

B-1-1）文境界認定

句点やその他の記号に基づき、文の境界(EOS)を認定し、その箇所に改行文字(¥n)を挿入する。この処理は単純な文字列置換であり、下記の文字の後に改行文字(¥n)を挿入する。

- ．　ピリオド
- 。　句点
- ）　半角カッコ閉
- （　全角カッコ閉
- ”　ダブルクォート
- ’　シングルクォート

この処理は、キーワードは文境界を越えて続かないと仮定し、長い文に対してキーワード候補の抽出数が多くなりすぎないようにするために行うものである。そのため、キーワードを分断することがない記号であれば、本来は文境界でなくても、境界と認定してしまって問題ない。

なお、本プロジェクトのOCR出力には、文章中の行送りに相当するTAB記号が含まれている。これはキーワード境界としても、文境界としても利用できないため、削除する。

B-1-2) 形態素解析

次に、各文を解析し、単語(形態素)列に分割する。今回は、オープンソースで公開されている形態素解析器Mecab[1]、及びIPA辞書 を用いた。形態素解析器は、文を形態素列に分割し、各形態素の属性(品詞や活用形など)を推定する。本処理で用いる形態素属性は、単語そのものと品詞のみである。なお、本報告書では形態素(morpheme)と単語(word)を特に区別せず、以降では「単語」というと「形態素」を指すこととする。たとえば、「リン脂質抗体症候群」という表現があったとすると、「リン」「脂質」「抗体」「症候」「群」という単語列に分解される。

B-1-3) 有限オートマトンによるキーワード候補抽出

前段の「リン脂質抗体症候群」という例では、「リン」「脂質」「抗体」「症候」「群」という単語列が得られた。これらの単語はすべて名詞であり、この中のすべての単語が専門用語(キーワード)である可能性がある。また、これらの単語を連結したもの(たとえば「リン脂質」「リン脂質抗体」「リン脂質抗体症候」「リン脂質抗体症候群」「脂質抗体」「脂質抗体症候」など)も、専門用語の可能性がある。

ここに挙げた専門用語候補のうち、「リン脂質抗体症候」「脂質抗体症候」は専門用語として不適当と思われるが、専門用語としての妥当性を一般的に決定できるルールは存在しない。また、これらの専門用語候補のうち、どれが文書を特徴づけるキーワードなのか、ランク付けを行う必要がある。

そこで、本プロジェクトでは単語列の品詞のパターンに注目して専門用語の候補を列挙し、それぞれの専門用語候補のキーワードらしさを統計処理によって推定することで、重み(重要度)付きキーワードリストを獲得する。本節(B-1-3)では、キーワード候補を列挙する手法を説明し、次節(B-1-4)では、統計処理によってキーワードの重み付けを行う手法を説明する。

本研究では、得られた単語列に対し、図2に示した決定性有限オートマトン(Deterministic Finite Automaton)を適用し、キーワード候補を抽出する。このオートマトンは、形態素列から以下の特徴を持つ表現(形態素部分列)を抽出する。

- 名詞の任意回数の連接をキーワード候補とする
- 1つ以上の名詞の列を助詞「の」をはさんで任意回数接続したものも候補とする

このDFAを状態遷移図として図示したものが図2である。なお、品詞体系は形態素抽出器によって異なることがあるため、DFAで抽出ルールを記述し、特定の規則をプログラムにハードコーディングすることを避けている。なお、DFAは文境界において必ず停止・再開する。

また、OCRの性質上、図などを文字と誤認識した場合、「禦撫灘塔鰹纒草笛製柵麺笹掻…」のような意味のない長い名詞列が入力されてしまうことがある。このような用語候補は、出現頻度が高くないため統計処理によって取り除くことができるが、無用に用語候補の数を増やし、処理速度の低下を招く。そこで、用語候補は最長で20語以内という制限を導入する。

[1] http://mecab.sourceforge.net/

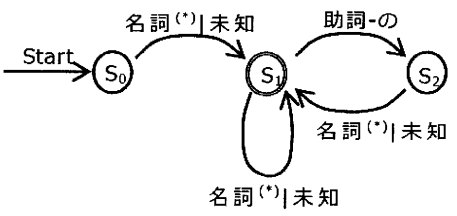
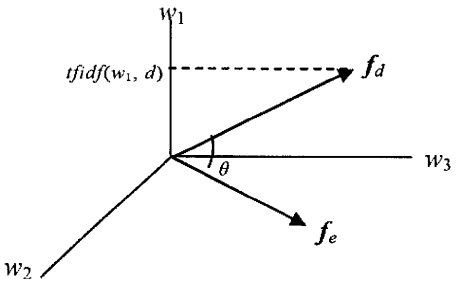


図 2 今回設計した DFA



- f_d を w_1 軸に射影した値が w_1 および d に対する tfidf 値である。
- f_d と f_e のなす角 θ は図示の通り。
- なお、実際には n 個の軸が存在する。

図 3 コサイン類似度の概念図

B-1-4) C-Value 値の計算

キーワード候補となる語列に対して、「キーワードらしさ」を推定するための統計処理を行う。本研究では、C-Value[2]と呼ばれる指標を採用した。この指標は、英語の生命・医学系の文献からの用語抽出として広く利用されてきた実績があり、日本語の文献に対しても適用可能であると判断した。

ある用語候補 s に対して、C-Value 値は次式で定義される。

$$\log_2(|s|+\alpha)\cdot\left\{n(s)-\frac{t(s)}{c(s)}\right\}$$

ここで、

- s はキーワード候補
- $|s|$ はキーワード候補の長さ(含まれる単語の数)
- α は短い語の重要度を調整する定数
- $n(s)$ は語列 s の出現数
- $t(s)$ は、s を包含する、s よりも長いキーワード候補の出現数
- $c(s)$は s を包含する、s よりも長いキーワード候補の異なり数

である。

「s を包含する、s よりも長い候補」とは、s を含むキーワード候補を表す。例えば s = 「脂質抗体」であったとき、「リン脂質抗体」や「リン脂質抗体症候群」のような語列を指す。

上述の定義式から明らかなように、ある候補語列 s の出現頻度が、その語列を前後に延長して得られる別の候補 S の頻度と同じである場合には、s は S 以外の語列には含まれないことになるため、$c(s)$=1 かつ $n(s)$ = $t(s)$であるから、s の C-Value 値は 0 となる。例えば、前記「脂質抗体」が仮に常に「リン脂質抗体」の形でしか現れなければ、「脂質抗体」の C-Value 値は 0 となる。逆に「脂質抗体」が多様なキーワードの一部として現れる場合には C-Value 値が大きくなるため、「脂質抗体」単独でキーワード候補となる可能性が高くなる。

α の値を制御することにより、1 語のみから成るキーワードのスコアを上げることができる(オリジナルの定義では α = -1 として、1語のキーワードを許可していない)。本研究では、実験的に α = 0.1 と設定した。

B-1-5) C-Value での用語候補の整列、上位候補抽出とストップワード

[2] Frantzi, K., Ananiadou, S. and Mima, H. (2000) Automatic recognition of multi-word terms. International Journal of Digital Libraries 3(2), pp.117-132.

前段では、キーワード候補とそのC-Value値のリストを得ることができた。キーワード候補をC-Value値の降順に整列し、C-Value値が1.0を下回らないという条件下で、C-Value値の高い用語候補上位10件までを研究報告書のキーワードとして出力する。

また、任意の語列をストップワードとして除外することができる。ストップワードは文字列のリストとして与えておき、キーワード候補がそのリストに含まれている場合には、上位であった場合でもその候補を採用しない。現状では、英語の一般語をキーワードから除くために使用している。

B-2) 類似キーワード・類似報告書の抽出

本節では、iGRANTSにおいてキーワード抽出に続けて行われる、「キーワード間類似度計算」及び「報告書間類似度計算」のアルゴリズムについて述べる。両者はいずれも、TF*IDFと呼ばれる特徴量を要素とする高次元特徴ベクトル同士のコサイン類似度を求める方式を採用している。まず、B-2-1節でTF*IDFについて説明し、B-2-2節で報告書間類似度、B-2-3でキーワード間類似度について説明する。

B-2-1) TF*IDFによる重み付け

TF*IDF値とは、文書に含まれる語が、その文書をどの程度特徴付けているかを表す統計的指標であり、任意の語 w と文書 d に対して、次のように計算される。

$$tfidf(w,d) = tf(w,d)\cdot idf(w),$$
$$tf(w,d) = \text{文書}d\text{における}w\text{の出現頻度},$$
$$idf(w) = \log_2 \frac{\text{総文書数}}{w\text{を含む文書数}}$$

tf (Term Frequency)は、語列 w が文書 d において何回出現するかを示す値である。*idf* (Inverse Document Frequency)は、文書群内で、w がその文書にどの程度集中して出現するかを表す値である。一般語は多くの文書に共通して表れるため、*idf* 値が小さくなる。両者を乗じることにより、w が特にその文書に顕著に表れる場合にTF*IDF値は大きくなる。

B-2-2) 報告書間類似度

すべての研究報告書に対して形態素解析(Mecab)を適用し、得られた名詞語のリストを $\{w_i \mid i = 1...n\}$ とする(すなわち、名詞の総数は n 個である)。ある研究報告書 d の内容を表すため、報告書dにおける名詞のTF*IDF値を要素とする n 次元ベクトルを作成する。

$$\vec{f}_d = (f_{d,1}, f_{d,2}, ..., f_{d,n})$$
$$= (tfidf(w_1,d)), tfidf(w_2,d), ..., tfidf(w_n,d))$$

2つの研究報告書 d と e が与えられたとき、これらの研究報告書の類似度は、それぞれの特徴ベクトル間のコサイン類似度(図3参照)として、次式で計算する。

$$\cos\left(\vec{f}_d \text{と} \vec{f}_e \text{のなす角度}\right) = \frac{\vec{f}_d \cdot \vec{f}_e}{|\vec{f}_d| \cdot |\vec{f}_e|}$$

報告書間類似度を求めるモジュールでは、あらかじめ $\vec{f}_d$ をすべての報告書に対して求めておき、任意の2つの報告書が与えられたとき、その類似度を上式によって即座に計算できるようになっている。

B-2-3) キーワード間類似度

キーワード間類似度の計算の原理は、報告書間類似度と同様である。異なるのは、報告書の特徴を表すベクトルを、キーワードの特徴を表すベクトルとして設計しなおす点である。キーワード間類似度では、各キーワードに対し、その出現箇所の前後 10 語(名詞のみ)から構成される文脈の和集合を作成し、各キーワードの特徴ベクトルとする。この特徴ベクトルは、各キーワードに対してどのような語が顕著に共起するかを表している。特徴ベクトルの各要素は、報告書間類似度と同様に TF*IDF 値を採用する。以上の方法によって特徴ベクトルを構成した後、類似度の計算手順は報告書間類似度と同様である。

C. 研究結果

厚生労働科学研究成果データベースに登録されている 1998 年から 2002 年度まで研究成果報告書(4634 件)に対し、知的検索のためのキーワード、類似キーワード、類似報告書を抽出する実験を行った。実験に用いた計算機は、Intel Core i3 540 (3.07GHz)、4GB DDR3 PC3-12800 RAM、Samsung HD040GJ HDD を搭載しており、64bit 版の CentOS 5.5 が動作している。知的検索に関するプログラムは、すべて Java で実装した。

4634 件の成果報告書に OCR 処理を行ったところ、約 7 億 4 千万文字(約 542MB)のテキストが得られた。B-1-1 節で説明した前処理を行ったところ、約 1900 万文が認識され、B-1-2 節の形態素解析器により、約 1 億 8 千万形態素に分割された。

B-1-3 節～B-1-5 節までのキーワード抽出処理を適用したところ、332 分(ディスク入出力時間を含む)を要して 19,776 個のキーワードが抽出された[3]。また、B-2 節の類似報告書・類似キーワード抽出に要した時間は、76 分(ディスク入出力時間を含む)であった。このとき、20,595,887 個のキーワード・ペア、10,708,796 個の報告書・ペアに対して、類似度の計算を行った。知的検索のためのインデックス処理は一度行えばよく、10 時間弱の処理時間は実用上問題にならない。

図 4 に「放射能」を検索クエリとし、本研究で開発された iGRANTS システムが返す検索結果を示した。検索された各報告書に対し、知的検索機能で付与されたキーワードが表示されている(赤枠で囲まれた箇所)。検索結果では、各報告書が利用者にとって有用かどうかを判断するため、報告書のタイトル、本文の要約(スニペット)、キーワードが表示されている。このうち、本文の要約は検索クエリの周辺の文を表示するものであり、報告書の本文全体を代表するような記述ではない。これに対し、キーワードは報告書の全体を代表するキーワードを 10 件まで提示する。利用者はこれらの情報を参照することで、検索された報告書が自分の検索要求に合致しているかどうか、迅速に判断できる。

図 5 は「テロ対策」というキーワードに関して、iGRANTS システムで類似するキーワードを表示させた時のスクリーンショットである。画面の左側では、「テロ対策」に関連するキーワードが、関連度の高い順に並んでいる。たとえば、「テロ対策」と関連の高いキーワードは「バイオテロ」関連度 74.15%)、「医師の救命」(関連度 70.04%)である。

図 5 のスクリーンショットの右側では、関連キーワードをグラフ形式で表示している。グラフ上の各点(ノード)が、キーワードで、各線(エッジ)は関連の深いキーワード同士を結ぶものである。「テロ対策」で関連キーワードを検索すると、「テロ対策」と関連の深いキーワード 10 件がグラフとし

[3] 各報告書より C-Value 値が高い上位 10 件のキーワードを抽出したときの、キーワードの種類数(異なり数)である。

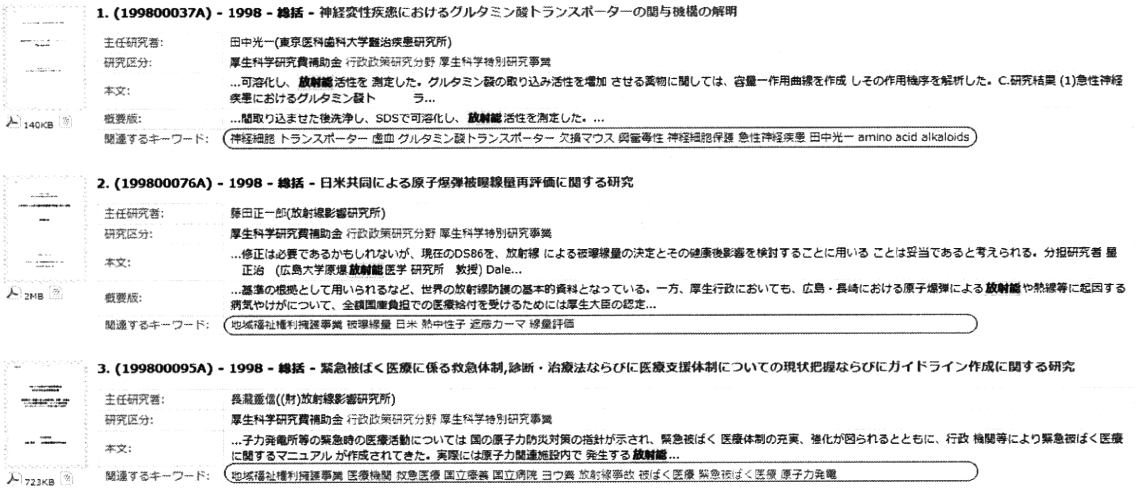


図４ 検索結果の画面に表示されるキーワード(赤枠で囲まれた箇所)

て表示される。ここで、「バイオテロ」というキーワードを右クリックし、「バイオテロ」の関連キーワードを表示させると、図５右のグラフが得られる。このように、入力したキーワードを出発点として、その関連キーワードをたどっていくことで、利用者が気づいていなかったキーワードにナビゲートすることができる。

D．考察

本研究で採用された手法を用い、知的検索のための情報(キーワードや類似度など)を生成するために要した時間は、数時間であった。この処理時間はインデックス化という観点から見れば、十分に高速である。知的検索のための処理では、研究キーワードの抽出に多くの時間を消費することが分かった。研究キーワードの抽出は、各報告書に対して独立に行う処理であるため、その処理時間は報告書数に対して線形に増加する。今後、インデックスされる報告書の数が増えたとしても、本研究の手法は十分にスケールすることが実証された。

本来、本研究で知的検索のために自動付与したキーワードは、研究を提案・実施した研究者自身によって与えられるべきものである。人手で付与されたキーワードがあれば、本研究の自動キーワード抽出の質を正確に評価したり、チューニングをすることで、本研究の成果を高めることができる。また、Medical Subject Headings(MeSH)のように、階層構造を持ったキーワード群を整備することができれば、類似キーワードの精度を高めることができる。

本研究でインデックス化された 1998 年から 2002 年までの成果報告書には、研究者自身によって付与されたキーワード情報が保管されていない。一方、文部科学省の科学研究費では、研究者自身による研究キーワードの登録・管理が行われており、報告書検索データベース「KAKEN」[4]で用いられている。文献データベースとしての価値を高めると共に、本ドメインにおける自然言語処理研究を発展させるため、報告書の本文のみならず、研究キーワードなどの付加情報の電子化は必須である。

本研究で自動獲得した情報により、キーワード間及び報告書間の類似性をグラフで視覚化した。本プロジェクトでは一般のユーザーが利用しやすい厚生労働科研成果デ

[4] http://kaken.nii.ac.jp/

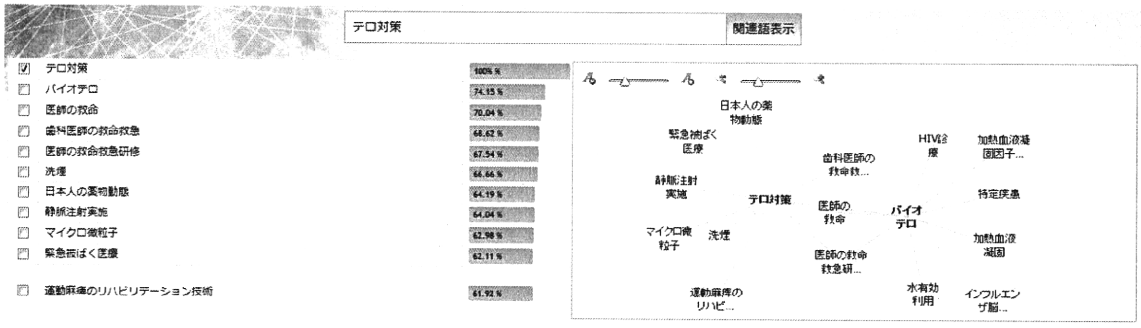


図５　「テロ対策」の関連キーワード

ータベースのあり方を考えるため、知的検索機能を盛り込み、iGRANTS システムを試験運用している。今後、実際の利用者からのフィードバックを収集することにより、研究成果報告書データベースがユーザーに積極的に提示すべき情報や、そのインタフェースに関して、検討が進められ、本運用のシステム設計に活用されることを期待している。

Ｅ．結語

本研究では、OCR 化された成果報告書に対する知的検索研究として、キーワードや類似キーワード・類似報告書の自動抽出の研究を行った。本研究の成果は、高機能版の厚生労働科学研究成果データベースのプロトタイプである iGRANTS に統合された。検索結果の画面では各報告書のキーワードが表示され、ユーザーは報告書の中身を即座に予測することが可能になった。類似キーワードや類似報告書を視覚化する機能により、ユーザーが探したい報告書へナビゲートする機能が追加された。

今後は、iGRANTS の暫定運用を通じ、実際のユーザーから知的検索機能のフィードバックを受け、機能の取捨選択を行う。必要な機能に応じて、人手により付与・管理すべき情報と、自動的に獲得すべき情報の切り分けを行うことで、成果報告書の電子化方法の検討が進むと期待される。

Ｆ．研究発表

１．論文発表

なし

２．学会発表

宇佐美佑, Han-Cheol Cho, 岡崎直観, 辻井潤一. 自動構築した大規模訓練データを用いた固有名抽出. 言語処理学会第17回年次大会(NLP2011), pp. 782—785, 豊橋技術科学大学(愛知県), 2011 年 3 月.

Han-Cheol Cho, Naoaki Okazaki, Jun'ichi Tsujii. Token Boundaries or Named Entity Boundaries. 言語処理学会第 17 回年次大会(NLP2011), pp. 778-781, 豊橋技術科学大学(愛知県), 2011 年 3 月.

Ｇ．知的財産権の出願・登録状況

なし

厚生労働科学研究費補助金（特別研究事業）
分担研究報告書

医学文献からの疾患情報自動抽出に向けた自然言語処理

研究分担者　建石 由佳（工学院大学情報学部）
奥村 貴史（国立保健医療科学院 研究情報センター）

研究協力者　野中 真生（工学院大学情報学部）
谷田 和章（東京大学学際情報学府）

研究要旨

研究成果データベースの検索機能を高度化に際しては、キーワードによる検索だけでなく、柔軟な質問に対応できる知的なシステムの構築という方向性がある。そのためには、既存の知識から情報を効率的に抽出していく自然言語処理技術が必要である。そこで、本研究分担では、厚生労働科学分野において重要性の高い疾患に関する知識ベースの整備を行い、また、この構築を自動化していくためのアノテーション環境について研究を行った。まず、遺伝疾患に関する医学文献を対象として、アノテータにより手作業での正解データを作成し、既存の情報抽出ツールの評価を行うとともに、性能向上の手法について検討を行った。また、アノテーション作業を効率化していくためのソフトウェアを検討し、汎用的なアノテーションツールの開発を行った。

A．研究目的

現在一般に利用されている Google などの検索エンジンでは、入力されたキーワードを文中から探し、キーワードと一致した文字列を含む文献を一覧表示する。インターネット上のこうした仕組みによって、情報検索の利便は飛躍的に向上した。しかし、指定したキーワードを含む文書を検索することは出来ても、欲しい情報そのものは、検索したユーザー自身が実際に文書の中から探し出す必要がある。

例えば、「血液型が変わることがあるか？」という疑問を持った際、「血液型」「変わる」というキーワードで検索をしても、血液型に関するページへのリンクと、「血液型」「変わる」を含む一部分が表示されるだけであり、疑問に対する答えを得ようとすれば、リンクを辿った先のページを確認しなければならない。また、適切なキーワードを思いつかない場合には、検索そのものが出来ない。例えば、発熱した子供の対応を検索しようとした際、「発熱」というキーワードを思いつかなければ、「体温が高い」や「熱が 38 度ある」という語句で検索をすることになるが、おそらく、望んでいる文書を得ることは出来ないだろう。

このような問題を解決するためには、例えば、「骨髄移植をすると血液型が骨髄提供者のものに変わる」という知識や、「人間の平熱は 36℃前後であり 38 度などの高熱は、“発熱”ないし“熱発”という異常な状

態である」という知識がなければならない。ところが、専門知識の多くは、人間が理解するための言葉(自然言語)で書かれたテキストとして存在しており、計算機が処理しやすいようなデータ構造を持ってはいない。

さまざまな知識を、計算機が処理しやすい構造にしたものを知識ベースといい、もっとも簡単な知識ベースとして用語集や辞書がある。しかし、医学のように高度に専門化した分野では、汎用的に利用できる知識ベースは存在せず、人間が読むことを前提にした医学辞典や機械処理に適した小規模な辞書はあるものの、知識ベースとしての利用が可能なリソースはほとんどないのである。

こうした状況を補うために、自然言語で記述されたテキスト自体から、計算機を用いて自動的に知識ベースを構築する研究が近年盛んになっている。こうした手法は、機械学習と呼ばれ、多くの場合、あらかじめ人間の専門家が抽出すべき正解がどこにあるかを示したテキストを与えることで、抽出の精度を向上させる処理を行うのが一般的である。このようなテキストの集まりを「コーパス」と呼び、この正解を示すためにテキストの該当する部分に正解であることを示したタグ(= 印)を付与する事を「アノテーション」と呼ぶ。アノテーションは、文献検索以外にも画像や動画などのコンテンツを検索するのにも用いられており、検索精度の向上を図るために、特に大規模な動画や画像の検索システムには不可欠なものとなっている。

しかし、特に医学のような高度に専門的な分野のテキストでは、疾患名、症状名、薬の名前など専門用語で書かれている場合が多く、専門家以外には何が抽出したい情報として適切なのかを判別するアノテーション作業を行うことは困難である。一方、正解を判別できるだけの専門知識を有する医学関係者は、アノテーションそのものの専門家ではなく、また、多忙であることも多いために、アノテーション作業の負担をできるだけ軽減する仕組みが必要である。

そこで本研究分担では、厚生労働科学研究成果データベースの検索機能強化に向けて、厚生労働科学研究分野において重要性の高い、疾患に関する知識ベースの構築とその構築を効率化に行うためのアノテーション手法の検討を行う。疾患知識ベースの構築としては、診断支援システム用に研究開発を進めていた疾患知識ベースを用い、臨床検査に関する文献より抽出した情報を追加していく手法について検討を行った。アノテーション手法の検討については、医学文献より疾患に付随する症状や理学所見などの情報を抽出する予備実験として、まず、アノテーターのトレーニングを兼ねた小規模なアノテーションと既存の情報抽出ツールを用いた情報抽出実験とを行った。また、こうしたコーパス作成作業を効率化していくために、汎用性が高く、直感的で簡単なアノテーションツールの開発を試みた。

B．研究方法

まず、疾患知識ベースの整備として、整備を進めて来た疾患知識ベースをクレンジングしたうえで、臨床検査に関する文献をテキスト化し、疾患知識ベース上の疾患コードとの対応を付けるという作業を試みた。たとえば、臨床検査Aの検査結果が高値だった際に疾患1、低値だった際に疾患2の可能性がある場合、症状マスタに「臨床検査A高値」「臨床検査A低値」というレコードを追加したうえで、疾患1の症状リストに「臨床検査A高値」を、また、疾患2の症状リストに「臨床検査B高値」を追加していく作業である。こうした作業を臨床検査項目の値毎に行う必要があり、膨大な作業量に上るため、効率化の手法についても合わせて検討を進めた。

また、医学文献からの情報抽出に際しては、まず、オンライン公開されている遺伝性疾患データベースOMIM[1]を対象として、小規模な手作業によるアノテーションと既存の情報抽出ツールを用いた自動的な情報抽出を試みた。OMIM においては、各遺伝性疾患の記述の中に、Clinical Featuresという疾患に関する情報が記述された節がある。ここには、その疾患の典型的な症状、検査所見、家族歴などが英文のフリーテキストで記述されている。

アノテーション作業としては、OMIM の Clinical Featuresに対して、疾患名、症状や所見に関する表現にそれぞれの意味クラスを正解タグとして付与したデータを作成した。後述するツールの作成が同時進行していたため、実作業は、XMLmindと呼ばれるXMLエディタを用いた手作業でのタグ付けと、Microsoft Word の蛍光ペンの機能を利用したタグ付けを併用した。

既存の情報抽出ツールを用いた予備実験としては、米国国立医学図書館(National Library of Medicine)で開発された概念タグ付けツールMetaMap[2]を利用し、上述の手作業によるアノテーション作業との比較を試みた。MetaMapは、文書を単語単位に分割し、シソーラスを用いて既知の概念と対応付けるツールである。たとえば、“Spinal imaging showed lumbar scoliosis”という文から“Spinal imaging”や“lumbar scoliosis”というフレーズを取り出し、さらに“lumbar scoliosis”を"lumbar”と“scoliosis”に分割したうえで、それぞれに[Body Location] と[Anatomical Abnormality]という生物医学概念タグを対応付けることが出来る。そこで、人手により作成した正解データとこの自動的な処理結果とを比較することで、既存ツールの性能評価を行った。

また、アノテーション作業を効率化するための手法についても検討を行った。そもそも、文章の中から、抽出すべき文言にタグをつけるという作業は、文献を読みながら重要箇所に線を引くという日常的に行われている作業に極めて近い。そこで、近年急速に広まってきたタブレット型 PC を用い、こうした「線引き」を模した作業環境を計算機上に実現することで、アノテーション作業に慣れていない医療従事者でも簡単にアノテーション作業を行っていけるツールの開発を試みた。

なお、本プロジェクトに限らず、医学関連のコーパスを作成する需要は少なくないことから、アノテーションツールには出来る限りの汎用性を持たせることが望ましい。そこで、ツールの作成に当たっては、汎用性の高いファイル形式である XML (Extended Markup Language)の枠組みを使用し、線引きにより実現するタグ付けの定義をユーザーが柔軟に行えるよう配慮した。

C．研究結果

C-1) 疾患知識ベースの整備

疾患知識ベースの整備としては、まず、構築してきた疾患知識ベースに対して、入力済みデータのクレンジング処理を行った。また、それぞれの疾患に関する情報を得るための参考文献リストを整理した。その上で、臨床検査に関する膨大な記述から、疾患と検査結果を紐付けるための作業を行った。

たとえば、「好酸球増加に対して寄生虫疾患の可能性」といった情報を機械処理可能な形へ変換するとする。その際、臨床検査や疾患に関する情報は、それぞれ自然言語で記載されているため、そのままコード化することが困難な記載が多数含まれていることが明らかとなった。たとえば、GOT上昇、といった情報に対して、「肝炎の可能

[1] http://www.ncbi.nlm.nih.gov/omim/
[2] http://metamap.nlm.nih.gov/

性」が示唆されている場合、「GOT 上昇」をコード化することは容易だが、「肝炎」のように、複数の疾患を包含した概念である場合、病名コードに一対一対応させることが困難である。

そこで、既存資料をどのような手順で処理すれば目的とする知識ベースが得られるのか検討を行った(参考資料 III-7)。結果的に、目的とするデータを手作業で作成する場合、かなりのコストを要することが明らかとなった。今後、後述するような自然言語処理技術を用いることにより、効率化を図っていく必要がある。

C-2) 疾患情報に対するアノテーション

予備実験として、医師 2 名、看護師 1 名により、OMIM に収録された 40 疾患に対してどのような症状が生じるかに関する記述についてタグ付けを行った。当初、症状に関する表現と合併症について抽出する方針でアノテーションを行ったが、疾患を特徴づける症状としてはこれらに留まらないさまざまな表現がなされていることが明らかとなった。

たとえば、病理学的な記載、病態生理学的な記載や、奇形に関する表現は、医学的には「症状」ではないが、確定診断に至るための重要な情報である。また、診断に役立つ臨床検査がある場合、あるいは、一般的な臨床検査でも特徴的な結果が生じる場合、これらも疾患知識ベースに収録すべき疾患の重要な“症状”データであると言える。さらに、疾患を特徴づける上では、陰性所見に関する文献上の記載も重要であり、実際、数多くの表現が認められた。その他、自然経過や好発年齢などについても、医学的には“症状”ではないものの疾患の表現系として重要な情報が数多く認められた。家族暦、合併症なども、同様の情報と考えることができるだろう。

また、希少疾患の場合、疾患の記述に、疾患に関する一般的な知識と個々の症例報告の情報が併記されることが少なくない。そうした場合、疾患データベースを対象として機械的な表現抽出を試みると、個々の症例の情報が過度に一般化されてしまう懸念が生じることも明らかとなった。

C-3) Metamap による情報抽出実験

次に、人的にアノテーションを行った20の医学文書に対して、MetaMapを用いることで機械的な情報抽出を行い、MetaMapの性能評価とチューニングを行った。評価としては、人手により症状情報として何らかのタグをつけられた部分がMetaMapによって[Finding] 、[Disease or Syndrome]、 [Sign or Symptom]などの疾患、症状に対応する概念タグをつけられている割合と、[Body location]+[function]のような概念タグの特定の組み合わせを症状表現の抽出事例として解釈した割合を測定した。MetaMapの出力に対する正否の解釈としては、人手で正解として何らかのタグをつけられた部分が明らかに疾患、症状に対応する概念タグをつけられた語を含むフレーズに含まれる場合を抽出成功とした。

実験の結果として、F値は53.3%であり、タグの特定の組み合わせを含むフレーズも抽出対象とした場合の抽出率は54.2%となった。なお、F 値とは、適合率と再現率の調和平均であり、2R/(N + C)で算出される。ここで、適合率はR/N、再現率はR/Cで表され、正解文書に含まれる正解タグの総数をC、実際に抽出された情報の総数をN、抽出した疾患に関する情報のうち正解であると判定されたデータの総数をRとしている。

抽出成功率が低い原因として、人手によるアノテーションでは対象となるフレーズを広く取ることが多い一方、MetaMapでは、かなり細かくタグ付けを行うことによる不

一致が頻発していたことが挙げられる。たとえば、人間は、“Collagen fibrils of the osteoid had a varying diameter”を症状情報として一体として解釈するが、MetaMapでは相互に独立したフレーズの集合に過ぎず、これらのギャップを埋める戦略を洗練させる必要があることが明らかとなった。また、数量表現の解釈や検査所見の処理など、フレーズの分類に留まらない文章の解釈を要する情報抽出において、MetaMapの課題が明らかとなった。

C-4) アノテーションツールの開発

最後に、指やペンを利用してラインマーカーで線を引くような感覚でタグ付けが可能となるアノテーションツールの開発を試みた。そのために、代表的なタブレット型PCであるiPad上に、アノテーションツールを実装した(参考資料III-6)。本システムを用いることで、アノテーターは、ラインマーカーを引くような直感的な操作によりタグ付け作業を行うことが可能となる。

操作としては、まず、タグをつけたいテキストとタグの定義ファイルを読み込む。タグの定義ファイルは、その文書中でどのようなタグや属性が使われているかを定義したXMLの文書型定義(DTD)の形式で記述されているため、極めて汎用性が高いものとなっている。タグとしては、複数のタグを登録することが可能であり、タグの種類ごとに異なった色を割り当てることが可能である。

タグ付けが終了すると、ラインマーカーが記入されたような色つきテキストを保存することが出来る。保存の際、文書はXML形式に自動的に変換され、ファイル全体がXMLファイルとして保存されると同時に、色とタグの対応付けを定義したCSSファイルが合わせて生成される。一旦保存したXMLファイルは、再度読み込むと前回の作業通りに色付けしたファイルとして表示される。また、保存時にできたCSSファイルを使用することにより、ブラウザなどでも確認することが可能である。

以上のアノテーションツールの開発により、従来はPCを利用した労働集約的な作業であったアノテーションが、持ち運び可能なタブレットPCを用いて他の業務の空き時間に作業を行うことが可能となった。今後、無償ソフトウェアとしての配布と、さらなる改良を計画している。

D. 考察

現在多くの検索システムで用いられるキーワードは、言語学的には名詞や名詞句にあたる。しかし、「血液型が変わることがあるか？」といった検索を行う場合、キーワードのような単純な仕組みでなく、質問文そのものを計算機が理解できるよう構造化することが１つの重要な研究課題であることがわかる。

そのためには、開発したツールを用いることでコーパスの分量を増やすとともに、タグ付けをされたフレーズを言語学的に分析し、係り受けの構造などをさらに詳しく調査しデータ化していくことで、コーパスの質を向上させていく必要がある。そのために、今回開発したアノテーションツールは、文章の重要な箇所にラインマーカーで線を引くという、日常的な作業に近い形でアノテーション作業を可能とするため、コーパス整備の効率を大きく改善するだろう。

ただし、現在のツールはプロトタイプであり、ツールそのものにも発展の余地が残されている。まず、今回のバージョンでは、単語間の掛かり受けを入力することが出来ない。また、色をメニューで選んでから線を引くという操作の煩雑性についても指摘されている。今後、使える色の一覧をパレットのように画面の一部に表示することで、

目的とする色を押しながらタグ付けしたい箇所に線を引くといった、より直感的なユーザーインターフェースの検討が必要である。

また、今回のツールでは、タグの定義をDTD形式で予め作成しておく必要がある。しかし、コーパス作成の初期の段階では、どのような情報をタグ付けすればよいかという点自体を探索的に決定していかなければならないことが少なくない。実際、今回のアノテーション作業でも、作業途中から家族歴をあらたにタグ付けすることになった。今後は、テキストと色付けから自動的にDTD形式のタグ定義を生成する仕組みなどを検討する必要がある。

E．結論

本研究分担では、厚生労働科学研究成果データベースの検索機能強化に向け、厚生労働科学研究分野において重要性の高い、疾患に関する知識ベースの構築とその構築を効率化に行うためのアノテーション手法の検討を行った。

疾患知識ベースの構築に際しては、構築を進めてきた知識ベースのクレンジングを行い、また、それぞれの疾患について記載した医学文献のリストを整備すると共に、臨床検査データの知識ベース化への追加について検討を行った。その結果、手作業での追加作業はコストが掛かることから、作業の効率化に向けた自動化の検討を進める必要性が示された。

次に、そのように医学に関する情報を自動的に抽出するために、機械学習の正解となるコーパスを効率的に作成するアノテーション手法の検討を行った。まず、手作業によりアノテーションを行い、正解データを作成した。その上で、既存の表現抽出ツールを用いて自動抽出した結果を評価し、改善に向けた課題を検討した。その結果、MetaMapのような自然言語処理システムが文を単語やフレーズに分割して認識するのに対して、専門家が「症状」として捕らえているものは、複数のフレーズにまたがったり、場合によっては文全体になるなどのケースが少なくないことが明らかとなった。さらに、正解情報となるコーパス作成に向けたタグ付けを簡便に行うためのアノテーションツールを検討し、iPad上にプロトタイプを実装した。

今後の課題として、医学文献の言語学的性質をより詳しく調べると共に、ツールの改良を進めることで、良質で大規模なコーパスを蓄積していくことが重要となるだろう。

F．研究発表

野中真生，奥村貴史，建石由佳，谷田和章，辻井潤一，「疾患プロファイル作成のための症状名抽出」，言語処理学会第17回年次大会(NLP2011)予稿集，pp. 643-646, 2011年3月.

G．知的財産権の出願・登録状況

1．特許取得

なし

2．実用新案登録

なし

3．その他

開発したXMLinemarker(アノテーションツール)を、無償ソフトウェアとして公開準備中。

厚生労働科学研究費補助金（特別研究事業）
分担研究報告書

クラウドを用いた仮設研究成果データベースの構築と運用

研究分担者　中村 修（慶應義塾大学 環境情報学部）

研究協力者　関谷 勇司（東京大学 情報基盤センター）
堀場 勝広（慶應義塾大学 大学院 政策・メディア研究科）

研究要旨

厚生労働科学研究では、従来運用されて来た研究成果報告書のメタデータ検索サービスであるGRANTSシステムに対して、全文検索を可能とする研究成果データベースの構築を目指している。そうした次世代GRANTSシステムにおいては、全文検索を可能とする性能と十分な耐障害性が求められる。

そこで、本研究分担では、次世代GRANTSシステムのような検索エンジンを、近年急速に普及しているクラウドコンピューティング環境上で構築、運用する実現可能性について検証を試みる。そのために、我が国を代表するインターネット研究コンソーシアムであるWIDEプロジェクトが運用するアカデミッククラウドであるWIDEクラウド上にiGRANTSシステムを構築し、仮設的なサービスの提供を行うと共に、クラウド環境でのデータベース提供に関する定量的な性能評価とサービス継続性に関する定性的な評価を行った。

結果として、構築したサービスが研究成果報告書の検索サービスに対して十分な性能を有していることを確認した。また、クラウド環境においてサービス構築を行うことで、ハードウェア障害やデータセンターそのものの機能障害が生じた状況においてもサービスを継続して提供しうることを示した。

A．研究目的

厚生労働科学研究では、従来運用されて来た研究成果報告書のメタデータ検索サービスであるGRANTSシステムに対して、検索機能を強化した次世代版の研究成果データベースの構築を目指している。一方、厚生労働科学研究は、毎年約1,500課題の研究を実施しているため、蓄積される報告書の数も膨大な量に上る。そのために、次世代の成果報告書データベースを提供する計算機システムも、十分な性能を有する必要がある。

また、このような公共サービスを行うシステムには、耐障害性と運用の継続性が要求される。たとえば、自然災害や局地的な停電によるサービス停止は、出来る限り避けられなければならない。実際、2011年3

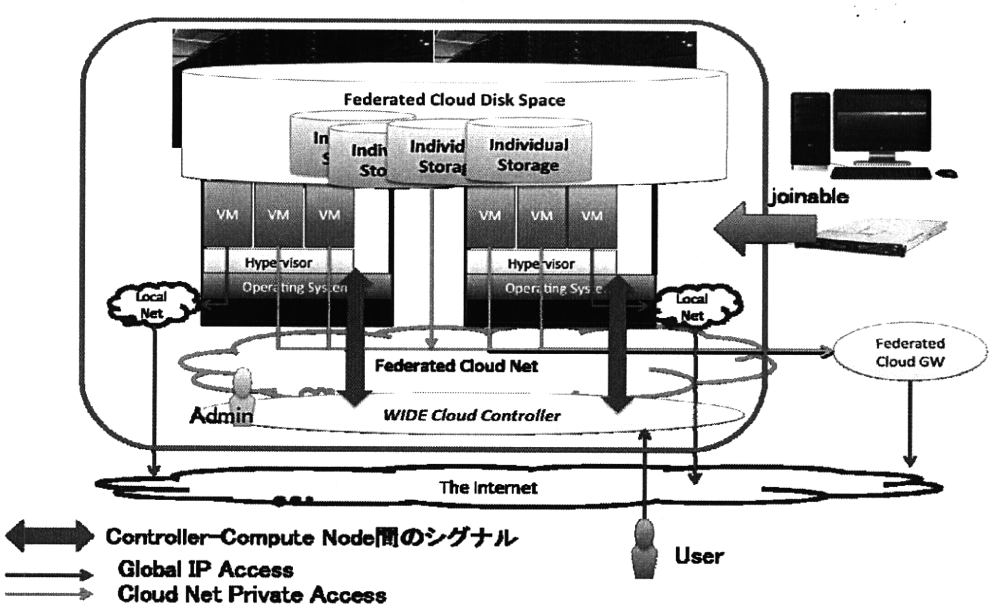


図 1 WIDE クラウド全体構成

月に生じた東日本大震災の影響によるデータセンタや通信網への被害は甚大であり、都心の一部地域を除いて首都圏の電力供給は極めて不安定となった。この事実は、情報システムの首都圏一極集中が有する危険性を再認識させ、緊急時の情報サービスの継続性確保について、再検証を求めている。

そこで、本研究分担では、次世代 GRANTS システムに必要とされる計算性能の確保とサービスの継続性を同時に満たすため、近年急速に普及しているクラウドコンピューティング環境上に検索エンジンシステムを構築することの実現可能性について検証する。具体的には、我が国を代表するインターネットの研究コンソーシアムである WIDE Project が運用するアカデミッククラウド「WIDE クラウド」を用い、次世代 GRANTS システムとして開発を進めている iGRANTS を構築する。また、その上で、構築したシステムの性能評価を行うとともに、サービス継続性に関する定性的な評価を行う。

B. WIDE クラウドの概要

WIDE クラウドは、参加する学術団体によって必要な計算資源を融通し合う連邦型のアカデミッククラウドであり、個々のユーザに独立した仮想マシンを提供する IaaS (Infrastructure as a Service)型のクラウドコンピューティングサービスである。この運営モデルにおいては、何らかの計算リソースを提供した学術組織は、他の組織が提供する計算リソースを利用できるようになる。

各研究組織は地理的に分散した拠点に存在するため、何らかの障害が特定拠点に発生した場合にも、他組織に設置された計算リソースが協調し、ユーザが構築した仮想マシンを継続して動作させることが可能である。そこで、この特徴を生かし、iGRANTS のような公共サービスに対して、クラウドコンピューティング環境上が必要な性能と継続性が提供できることを検証する。