Science Journal of Kanagawa University **29** ：97-99 (2018)

■報告書■　2017 年度神奈川大学総合理学研究所共同研究助成論文

医学用語の統合によるカルテの潜在意味解析の研究

韓　　浩[1, 3, 4]　中山　堯[2, 3]

Semantic Analysis of EHR Based on Medical Terminology

Hao Han[1, 3, 4] and Takashi Nakayama[2, 3]

[1] Department of Healthcare Information Management, The University of Tokyo, Bunkyo-word, Tokyo 113-8655, Japan
[2] Department of Information Science, Faculty of Science, Kanagawa University, Hiratsuka City, Kanagawa 259-1293, Japan
[3] Research Institute for Integrated Science, Kanagawa University, Hiratsuka City, Kanagawa 259-1293, Japan
[4] To whom correspondence should be addressed. E-mail: han@hcc.h.u-tokyo.ac.j

Abstract: This research establishes and promotes a project that aims to collect and arrange medical term dictionaries to facilitate NLP in clinical settings.
Keywords: EHR, medical term

序論

診療記録テキスト検索システムの利用者（医療従事者､医学研究者等）は１つのデータに対してすら様々な検索意図を有するため、利用者の満足度を高めるためには多様な検索意図と診療情報の意味関係を反映させて検索結果を解釈し、それを利用者に分かりやすい形で出力する必要がある。このために、本研究では、①潜在的な検索意図の分析、②検索意図と情報の意味関係を反映した出力方法、及び①の結果から②を生成するための③検索モデルとアルゴリズムの３つの課題を、相互に密接に連携する要素として位置づけた上で、それぞれの課題の解を与えることにより、一貫性のあるシステムを構築可能とすることを目的とする。2017 年度は検索処理について実現に必要なデータを明らかにし、情報の意味関係の認識のため医学統合辞書検索の構築を行った。

研究背景

SS-MIX2[1] 拡張ストレージ文書検索エンジンを代表とする現在の診療記録テキスト検索システムは、利用者の多様化する情報要求に十分に応えられているとは言えない。その要因として以下のような問題がある。

利用者が入力する問合せ（クエリ）は情報要求を的確に表現していない場合が多く、特に入力されたキーワードが少ない（1-3 個）場合はクエリが曖昧あるいは不十分になることが多い。これには、多義語のように検索対象自体に複数の解釈が可能である場合や、検索対象が様々な側面を有する場合（例えば、肺癌ではレントゲンの所見、治療計画、喫煙歴等）などがある。ところが、従来の診療記録テキスト検索エンジンでは、キーワードを含む文書（診療記録テキストなど）を、クエリと文書との類似性（relevance）のスコアに基づいてランキングするため、検索結果の上位が特定の解釈や側面に関する文書で占められるという状況が避けられない。この結果、少数派の多様な情報要求にはほとんど応えられないという問題が生ずることになる。

具体的には、検索意図マイニングにより、多義性のあるクエリ「○○」の潜在的な検索意図として「○○ ××」が得られたとしても、目的の文書には必ずしも「××」（及びそれに類似のキーワード）が含まれているとは限らないため、このような文書は「○○ ××」に対する検索結果として得ることはできない。これは再現率を低下させる問題の一つである。例えば、クエリ

「胸部 レントゲン」

の検索意図が

「胸部レントゲン検査結果を調べる」

が得られたとしても、現存の診療記録テキスト検索エンジンでは

「Xp 上アレセンサが効いている」
のような検索結果が得られない。それは、検索システムに「レントゲン」と「Xp」（Xp はレントゲンの略語）あるいは「アレセンサ」と「胸部」の意味関係（アレセンサは肺癌を治療する薬物である）を認識する機能が実装されていないことが原因となっている。

逆に、検索結果になる文書に対して、検索キーワード「○○ ××」が含まれていてもそれが検索意図を反映していない場合もある。これは適合率を低下させる問題の一つである。例えば、クエリ
「妊娠 高血圧」
に対しては、次のような検索結果が得られる可能性がある：
「本人：不妊治療 父：胃癌 母：高血圧 姉：妊娠中」
「妊娠中の高血圧特になし」
しかし、これらはいずれも実際の検索意図
「妊娠 高血圧症患者を調べる」
に合わない可能性がある。

上記のような問題点は、従来の手法が主として文書に表現された字面の情報を対象として研究されてきたことを示唆している。すなわち、利用者の検索意図と診療情報の意味関係を考慮することなく、文書の分類結果に基づいて情報の意味を判断するという分析の客観性を欠いているという根本的問題がある。

具体的には、これまでの国内外の研究は、検索対象になる文書のクラスタ解析に基づく利用者の閲覧行動の分類分析を中心として行われている。この手法は一部の検索エンジンでも重複・類似文書を除外する技術等として実用化されている。また、検索意図マイニングや類似の研究も従来から鋭意進められており、主に検索ログやクリックスルーログの分析に基づく技術としてクエリ補完やクエリサジェストの機能が実用化されている。一方、検索結果を分類あるいはクラスタリング（＝文書内容による組織化）して出力する研究も診療記録検索エンジンの出現後比較的早い時期から行われてきており、実用化されたものもある。しかし、このように個別の課題設定においては研究や実用化が進展しているものの、これらは単独では上述した問題への十分な解とはならない。検索意図の一連の変化の理解と検索満足度の分析を行わず、主観的な判断に依存していることを避けられない。

計画

利用者の理解しやすさと利便性を考慮して、検索結果の表示出力において取り扱うべき検索意図を確定する。ここでは、検索意図と検索満足度に人手で付与している属性を参考として、潜在的な検索意図を抽出して列挙する方法を研究する。その際、検索意図に関連がありそうな文書に着目することにより高精度化を図る。そのために、列挙した潜在的な検索意図に対して、種別と確率を付与する方法を検討する。種別については以前の研究[2]に付与された検索意図種別を学習データとし、機械学習を利用して付与することを想定している。確率については各種の情報資源中における検索との共起頻度等に基づく推定手法を開発し、被験者実験により評価を行う予定である。

多義性等のあるクエリに対して、多面性を有する対象文書に対する情報の意味関係と利用者の検索「コンテクスト」などの扱いについて検討を行う。そして、既存の医学用語リソース等を利用して、さらに所属研究機関に開発された医療知識基盤としての臨床医学オントロジーを導入して、自然言語処理の方法に基づいて診療記録の意味関係を分析することにより情報の概念・属性と繋がり関係等を取得する。

実装

医学用語の総合検索サービスを提供するため以下の医学辞書を整理・統合して Web サービス API を開発している。

- Comejisyo
- ICD-10
- LiLak
- UMLS
- 標準病名マスター
- 医薬同義 T 辞書
- 医薬品 HOT コードマスター
- JAPIC 薬剤データベース
- 臨床検査マスター

開発環境として以下のものを利用している：

- Server Version: Apache Tomcat 8.5.30
- JVM Version: 1.8.0_60
- Jersey[3] Version: 2.25.1

開発済みの API の例をいくつか以下に示す。

API1：単語があれば、単語のコンセプト ID と存在先を返す関数 term2attributes
例：「イスポール」の属性を調べる場合：
http://localhost:8080/meddic/term2attributes?term= イスポール
#term：パラメータ名、イスポール：パラメータ値

検索結果は、以下のような JSON のフォーマットとして返す。これによって「イスポール」は医薬同

義辞書の default 表（コンセプト ID:D00002）に含まれていることが分かる。


```
{
 "term": " イスポール ",
 "id_attributes": [
  {
   " 医薬同義 .default": {
    "concept_id": "D00002"
   }
  }
 ]
}
```


API2：単語コンセプト ID で指定された辞書から単語の属性を返す関数 conceptID2attributes

例：医薬同義辞書からコンセプト ID(D00002) の対応する単語の属性を調べる

http://localhost:8080/meddic/conceptID2attributes?composite_resource_name= 医　薬　同義 .default&　concept_id=D00002

composite_resource_name：辞書名、concept_id：コンセプト ID

結果は、以下のような JSON のフォーマットとして返す。これによって医薬同義辞書の中における（ID:D00002）の単語に関連する属性（上位関係、カテゴリ、同義語リスト等）が分かる。


```
{
 "composite_resource_name": " 医薬同
義 .default",
 "concept_id": "D00002",
 "composite_standard_codes": [],
 "parents": [],
 "children": [],
 "categories": [
  " 医薬品名 "
 ],
 "lead_term": 1,
 "synonym_set": [
  "12% 総合アミノ酸製剤 ",
  " イスポール ",
  " ヌトリゾール ",
  "12% MIXED AMINO ACID
PREPARATION",
  "ISPOL",
  "NUTRISOL"
 ],
 "inferred_categories": [
  " 医薬品名 "
 ],
 "others": []
}
```


今後の予定

開発された医療知識基盤としての臨床医学オントロジーを導入して、自然言語処理の方法に基づいて診療記録の意味関係を分析することにより情報の概念・属性と繋がり関係等を取得する。実験を行って基本的な有効性の確認を行うとともに、Wikipedia 等により収集したデータも考慮して意味情報の追加導入を行う。また、開発した手法を効率的に実行するために並列処理を導入して実装する。

謝辞

本研究は、研究課題「検索クリエ分析によるカルテの潜在意味解析の研究」に対する 2017 年度神奈川大学総合理学研究所共同研究の研究助成（RIIS201704）を受けて行いました。厚く御礼申し上げます。

文献

1) SS-MIX2. [http://www.ss-mix.org/cons/ssmix2_about.html].
2) 韓　　浩，郭　俊霞，中山　堯 (2015) Q&A コミュニティに注目したブラウジング行動に基づく検索満足度の予測分析．*Sci. J. Kanagawa Univ.* **26**: 41-45.
3) Jersey. [http://jersey.github.io/].