5R-9

非専門家による使用を想定した医療情報処理システム

佐藤 敏紀 † 上笹 正典 ‡ 三浦 弘之 ‡ 渡辺 昌貴 ‡ 上原 貴夫 §
東京工科大学大学院 † 東京工科大学工学部 ‡ 東京工科大学コンピュータサイエンス学部 §

1 はじめに

近年，インフォームド・コンセント (Informed Consent) や EBM(Evidencebased Medicine) を実践する医師が急速に増えている．そのため医師および医療関係者が他者に提供する医療情報の量が増えている．それと同時にインターネットが一般家庭に広く普及したことで，医師および医療関係者や企業が発信する医療情報を誰もが気軽に入手できるようになった．

医療情報についてインターネットを利用し調査する際には，検索クエリとして医学用語を的確にもちい，入手できる膨大な情報から正確かつ必要な情報のみを抽出しなければならない．しかし非専門家はこれらの行動を理想的におこなえないケースが多い．また近年，Weblog が旧来の日記サイトと同じ用途で普及しており，Google や Yahoo!!など主要な検索エンジンの検索結果には顕著なノイズが混入するようになった．以上により，情報検索により誤った根拠を入手し医療情報を誤って解釈する可能性が高まっている．

一方，日本国内における医学用語は分野や書籍により表記や邦訳が異なることが多く，医学用語の非統一性は誤った医療情報の発生源になるため問題である．しかし日本医学会など医学用語の統一を目指している団体が大きな成果があげるまでには時間がかかる．

本稿では上記の問題を少しでも解決するために医療情報処理システムを提案する．我々が提案するシステムにより，非専門家が医療情報群から最適な情報を検索エンジンをもちいた単純な検索と比べ，低リスクかつ低コストで抽出できる．

2 医療情報を扱うために

我々のシステムは医療情報を扱うため，他の研究と比べ前提とする知識が多い．そこで医療情報を扱うために必要な知識や資源について説明する．

2.1 ユーザの分類

本研究では我々が構築するシステムのユーザとなりうる人々を分析した．医師および医療関係者ではない一般人に対しては大規模なアンケートを実施して [1]，彼らが医学用語として使う言葉に関して調査をした．その結果に基づき，すべてのユーザは「専門家かつ医師および医療関係者」,「非専門家かつ医師および医療関係者」と「非専門家かつ一般人」の 3 種類に大別できる.「非専門家かつ医師および医療関係者」は自分の専門分野外の医療情報に対して的確な判断が下せるとは限らない.「非専門家かつ一般人」は医学用語のつもりで一般語 (医学用語としては不適当，不正確な用語) をもちいることが多い．一般語については先述のアンケートにより，一般語がクエリとして不適当であることを確認した．

Medical Information Processing System for Non-professional users

† Toshinori Satoh,Tokyo University of Technology
‡ Hiroyuki Miura,Masanori Kamisasa,Masaki Watanabe,Tokyo University of Technology
§ Takao Uehara,Tokyo University of Technology

2.2 医学用語辞書の作成

検索エンジンで医療情報を検索する際にユーザはキーワードとして医学用語を入力しようとするが，専門家でも医学用語の表記や使用法を間違える場合が多く，一般人は医学用語のつもりで一般語を入力することが多い．また，ユーザが検索に最適なキーワードを入力できるとは限らない．よって我々はユーザが使用したキーワードを解析し拡張すべきだと考え，そのために医学用語辞書を作成することにした．

医学用語辞書用の辞書データは IPA 品詞体系に [2] 従って記述する．本研究において必要なデータが IPA 品詞体系で未定義の場合は，IPA 品詞体系の形態素エントリを互換性を保ちながら拡張する．その際，他の研究との資源共有が容易であるように配慮する．作成する医学用語辞書には医学用語の意味を記述せず，以下のように 2 要素または 3 要素を 1 エントリとして扱う．エントリには品詞情報や読みなども記述する．

- 英 (欧) 語表記の医学用語，日本語表記の医学用語
- 英語表記の医学略語，英 (欧) 語表記の医学用語，日本語表記の医学用語
- 対応する医学用語がある一般語，日本語表記の医学用語，正確な英 (欧) 語表記の医学用語

本研究で作成する医学用語辞書は登録するエントリ数が 10 万エントリ以上である．よって，エントリ内の各要素に設定する必要がある形態素生起コストを手作業で決定することは，膨大な作業量と医学用語の専門性から避けるである．そこで作成する医学用語辞書の形態素生起コストは医療情報をコーパス化したものなどから自動取得するべきであると考えた．本研究では機械学習による形態素生起コストの自動取得を実現する．初期段階は HMM(Hidden Markov Model:隠れマルコフモデル) による学習を実装する．形態素生起コストを学習する際に使用するコーパスの１つとして，CSJ(日本語話し言葉コーパス:the Corpus of SpontaneousJapanese)[3] を検討している．CSJ は UniDic[4] の形態素生起コスト算出にもちいられている．一般人は情報検索時に口語に近いクエリを入力することが多いため，CSJ による学習は一般語や日本語話し言葉について良い解析結果が期待できる．

本研究において作成する医学用語辞書には日本語の医学用語が収録されるため，異表記同語問題 [5] が発生する．本研究では異表記同語問題への対応を一貫しておこなうため，他の辞書作成に関するガイドラインの内容をふまえ独自のガイドラインを作成し，従うことで日英対訳，異表記同語への対応，一般語から医学用語への変換などを見据えた辞書データを作成する．

本研究において作成する医学用語辞書の登録作業は人手でおこなっている．med-ipadic は作成をはじめてまだ間もない辞書であるため，収録しているエントリ数が十分ではない．そこで作成する医学用語辞書と ipadic や MeSH などの資源を連係させて不足しているエントリを補いたいと考えている．MeSH とは NLM(National Liblary of Medicine:アメリカ国立医学図書館) が作成している XML で記述されたシソーラスである．MeSH は最低でも年に

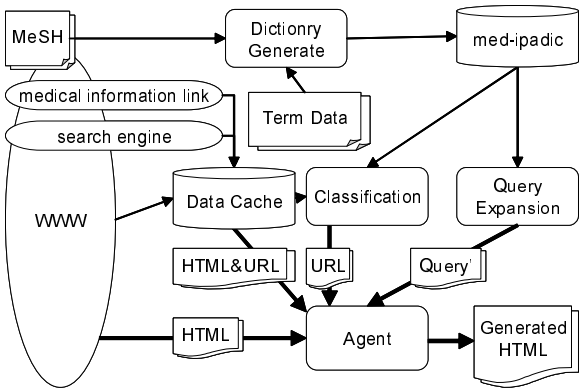

図 1　構築する医療情報処理システムの概念図

2 回更新される．MeSH には日本語が記述されていないため，日本語への翻訳は手作業でおこなうことになると思われる．

3　システムの実装

我々が構築するシステムの概念図を (図 1) に示す．

このシステムはユーザがキーワードを入力する以前に Dictionary Generate モジュールによって，医学用語辞書を記述した CSV 形式のデータや MeSH などを用いて IPA 品詞体系に従い記述された med-ipadic(医学用語辞書) を生成する．Agent を含む多くのモジュールが med-ipadic を参照し動作する．

このシステムの動作はユーザが Agent にキーワードを入力することではじまる．ユーザが入力したキーワードは Query Expansion モジュールにおいて，形態素解析などの自然言語処理を経て拡張される．Agent は拡張されたクエリである Query' をもちいて，ユーザの指示に従い動作する．Query Expansion モジュールはユーザの許可が得られた場合，med-ipadic の記述に基づきクエリを修正する．よって，ユーザが複雑な医学用語を入力した場合に発生する誤字脱字や，一般人が入力する一般語や日本語話し言葉をより最適なクエリに変換できる．

Agent は単純な検索を指示された場合は，指定された Web サイトから医療情報を収集し結果をマイニングして HTML ファイルを生成する．その際 Agent はあらかじめ手作業で選別された，安全な情報が得られる可能性が高い URL のリストから収集先を決定する．

Classification モジュールは Agent の指示に従い Web サイトを巡回し，Agent に指示された件数だけ得た医療情報を Query' にもとづきテキスト分類することで，Query' に最も近いと考えられる順に並べ替えた URL リストを得られる．Classification モジュールは得た URL リストを Agent に返す．よって Agent は URL リストにしたがい Data Cache またはインターネットから医療情報を収集することで，よりユーザの意図に近い検索結果を生成することができる．分類アルゴリズムとしては主成分分析，独立成分分析，Support Vector Machine を実装した．ユーザはキーワード入力時に任意の分類アルゴリズムを選択することができる．

4　実験，評価について

本研究で構築したシステムを評価するため，我々のシステムを実際に 1 週間運用して得られた医療情報により評価をおこなう．ユーザが我々のシステムに入力したキーワードを回収し，そのキーワードをもちいて一般に普及している検索エンジンから医療情報を得る．その情報と，同じキーワードによって我々のシステムが収集した医療情報を比較し，キーワードに対する解答の精度や，入手した医療情報としての質から評価をおこなう．我々のシステムに含まれる各モジュールについても，それぞれ妥当な評価をおこなう．

また非専門家による利用を想定し，既存の情報検索システムでは解決できない質問セットをもちいた実験をおこなう．質問はクエリとユーザの目的を一組とし情報検索結果とユーザの目的を比較し，ユーザの目的を達成できた件数により我々のシステムを評価する．

5　おわりに

本稿では非専門家による利用を想定した医療情報検索システムを提案した．一般語や日本語話し言葉に対応した辞書を作成し，それをもちいた医療情報検索について述べた．また実験方法や評価方法について述べた．我々のシステムのように専門情報を少しでも噛み砕く仕組みは需要が多い。

本研究の課題は複数あげられる．辞書に関してはフォーマットを UniDic 品詞体系 (XML 表記) に変更することで，XML で記述されている MeSH から得られるデータを有効に活用できる．また MeSH は日本語訳が無いため，日本語の医学用語に対応するため効率の良い翻訳手法を模索する必要がある．翻訳は，すべての作業を自動にできないとしても自動化された割合を増やすことが重要である．本研究では形態素解析器と文書分類器を自作しているが，その性能は先行研究 [6] と比較して劣る．これらのツールの性能を向上させることでシステム全体の性能も向上するであろう．

今後の展開として，システムの継続的な運用や信用に基づく情報を利用した医療情報源の確保がとても興味深いと考えている．

参考文献

[1] 佐藤 敏紀, 上原 貴夫, エージェントによる医療情報検索 - 一般人向け医学用語辞書の実装 -, 第 66 回情報処理学会全国大会,2U-8,2004

[2] 浅原 正幸, 松本 裕治, ipadic version 2.7.0 ユーザーズマニュアル, 奈良先端科学技術大学院 大学情報科学研究科 自然言語処理講座, 2003

[3] 前川 喜久雄,「日本語話し言葉コーパス」の概観, 国立国語研究所, 2004

[4] 伝 康晴, 宇津呂 武仁, 山田 篤, 浅原 正幸, 松本 裕治, 話し言葉研究に適した電子化辞書の設計, 第 2 回「話し言葉の科学と工学」ワークショップ, pp.39-46, 2002

[5] 佐藤 理史, 異表記同語認定のための辞書編纂, 情報処理学会研究報告, 2004-NL-161 (14), 2004

[6] 松本 裕治, 北内 啓, 山下 達雄, 平野 善隆, 松田 寛, 高岡 一馬, 浅原 正幸, 日本語形態素解析システム『茶筌』 version 2.2.1 使用説明書, 奈良先端科学技術大学院 大学情報科学研究科 自然言語処理講座, 2000