E-011

非専門家による日本語話し言葉をもちいた医療情報検索
Medical Information Retrieval in Spontaneous Japanese which non-professional uses

佐藤 敏紀 † Toshinori Satoh
上原 貴夫 ‡ Takao Uehara

1 はじめに

近年、インフォームドコンセント (Informed Consent：医師による説明とそれに対する患者の同意) や EBM(Evidencebased Medicine：根拠にもとづく医療) 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 Medical Information(医学，医療分野の情報) の量も増えている．Medical Information は専門性が高いため，これを入手した場合はインターネットや書籍をもちいて積極的に意味を深く理解するべきである．Medical Information についてインターネットを利用し調査する際には，検索キーワードとして Medical Terms(医学，医療に関する用語) 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するべきである．しかし非専門家は的確な検索キーワードを入力できないことや，入手した情報から正確かつ必要な情報を抽出できないことが多い．そのため近年，情報検索により誤った情報を入手し Medical Information を誤って解釈するケースが増えている．

一方，日本国内における Medical Terms は分野や書籍により表記や邦訳が異なることが多く，Medical Terms の非統一性は，誤った Medical Information の発生原因として問題視されている．しかし企業や大学病院としては用語統一により得られる利益が少ないため問題解決に対する取組は消極的である．また日本医学会をはじめ日本国内の Medical Terms を統一しようという活動をしている人物や団体が，大きな成果があげるまでには時間がかかる．よって，非専門家であろうとも専門性の高い情報群から最適な情報を低リスクかつ低コストで抽出できるシステムが必要とされている．

本研究の目的は非専門家による正確かつ的確な医療情報検索を実現することである．本稿では非専門家による日本語話し言葉をもちいた医療情報検索を実現するための手法として以下，本研究の前提条件である 2 章にてユーザ分類について述べる．3 章にて，医療情報検索のためにおこなう自然言語処理でもちいる医療，医学関連用語辞書や形態素解析器の作成，実装方法について述べる．4 章にて作成する辞書の評価方法について述べる．

2 ユーザの分類

本研究ではユーザの分類基準として、医師および医療関係者と一般人、専門家と非専門家の 2 種類を定義する．

2.1 医師および医療関係者と一般人

ユーザが Medical Terms を正しく使うことができるかどうかに着目し，医師と Medical Terms を正確に扱える医療従事経験者を医師および医療関係者に，医師および医療関係者以外は一般人に分類する．医師以外の人物は Medical Terms を正確に扱えなければ現役で医療に従事しているとしても一般人として扱う．

2.2 専門家と非専門家

ユーザがある瞬間に扱う情報の分野がユーザの専門分野かどうかに着目し，自らの専門分野の情報を扱う人物を専門家に，専門家以外は非専門家に分類する．日常において多くの一般人には医師および医療関係者がすべて医学，医療の専門家に感じられる．しかし現実には専門領域外の患者を診療している医療機関も多く，この分類基準を意識していない一般人が非専門医の診療を受け誤診されるケースが頻発している．

医学，医療の分野において非専門家は，医師および医療関係者かつ非専門家と一般人かつ非専門家に分類できる．

3 医療医学関連用語辞書

3.1 辞書のデータ構造

本研究では医療，医学関連用語辞書を作成する．辞書データの記述は，ほぼ IPA 品詞体系に [8] 従う．本研究において独自に必要なデータは，IPA 品詞体系との互換性を保ちながら拡張することで，他者がおこなう研究との資源共有を容易にする．作成する辞書は用語の表記や性質に着目しており用語の意味は記述しない．用語を収録する際には以下のような 2 語または 3 語を 1 組とし扱う．例として IPA 品詞体系にもとづき記述した辞書データを図 1 に示す．

- 英字表記の Medical Terms，日本語表記の Medical Terms
- 英字表記の医学医療関連略語，英 (欧) 語表記の Medical Terms，日本語表記の Medical Terms
- 対応する Medical Terms がある一般語，日本語表記の Medical Terms，正確な英 (欧) 語表記の Medical Terms

```
(品詞 (名詞 一般 医学))((見出し語 (高病原性鳥インフルエンザ 4649))(読みコウビョウゲンセイトリインフルエンザ)(発音 コウビョウゲンセイトリインフルエンザ)
(複合語
((品詞 (接頭詞 名詞接続))(見出し語 高)(読み コウ))
((品詞 (名詞 一般))(見出し語 病原)(読み ビョウゲン))
((品詞 (名詞 接尾 一般))(見出し語 性)(読み セイ))
((品詞 (名詞 一般))(見出し語 鳥)(読み トリ))
((品詞 (名詞 一般 医学))(見出し語 インフルエンザ)(読み インフルエンザ))
)),Highly Pathogenic Avian Influenza
```

図 1 IPA 品詞体系にもとづく辞書データ例

3.2 形態素生起コストの算出法

本研究で作成する辞書には IPA 品詞体系と互換性をもつために，形態素生起コストを登録する必要がある．この形態素生起コストの設定方法として，ipadic に登録されている形態素生起コストを参考にし適時微調整するという方法がある．奈良先端大学院大学の松本研究室が作製した形態素解析システム「茶筅」[7] は，IPA 品詞体系をもとに作成した ipadic(IPA 品詞体系辞書) を利用している．茶筅が普及するにつれ IPA 品詞体系と ipadic も普及し，数多くの日本語処理ソフトウェアで利用されている．しかし，本研究の場合は登録する品詞数が膨大であること，検索対象が専門性の高い医学用語であることから，

† 東京工科大学大学院,Tokyo University of Technology
‡ 東京工科大学コンピュータサイエンス学部,Tokyo University of Technology

形態素生起コストは医療情報をコーパス化したものから自動で取得することが望ましい．そこで本研究では独自に形態素解析器を作製し，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) による学習と viterbi アルゴリズムをもちいたコスト最小法をもちいて実装し [2][3]，先行研究 [6] を参考に順次改良する．

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，日本語話し言葉コーパス (the Corpus of Spontaneous Japanese)[9] を検討している．日本語話し言葉コーパスは格納されている音声ファイルの素性から，一般語を多く含んでいるとは考えられない．しかし，一般人が情報検索時に話し言葉に近い検索キーワードをもちいるケースが多いことから，未知の形態素定義を抽出することや，形態素生成コストの学習することにもちいることで，より良い実験結果が得られるのではないかと考えている．

3.3 非専門家への対応

本研究では医師および医療関係者が使用する範囲の Medical Terms および略語，隠語，俗語を正しい Medical Terms とし，一般人がもちいる正しい Medical Terms とはいえない用語をすべて一般語として定義した [1]．非専門家への対応として，非専門家により検索キーワードとして入力すされる一般語を収集し，正しい Medical Terms と対応させることで非専門家の情報検索を補助する．

3.4 異表記同語問題への対応

本研究において作成する辞書には日本語の Medical Terms が収録されるため，異表記同語問題 [5] が発生する．これは Medical Terms の非統一性とも関連する問題である．辞書を作成する際に参照する資料が複数存在し，そのほとんどが表記の統一をおこなっていない．また 1 つの用語について複数の表現を掲載していることもある．各資料を同じ意味の用語について比較すると表記が統一されていない場合が多い．このような異表記同語問題への対応を一貫しておこなうため，他の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する．

3.5 作成する辞書以外の資源との連係

作成する辞書で対応できなかった語については，作成する辞書と ipadic や MeSH などの資源を連係するインターフェイスにより補うことができると考えている．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 により記述たシソーラスである．MeSH は最低でも年に 2 回更新される．ipadic の品詞収録数や MeSH から得られる国際標準の最新情報は魅力的である．

3.6 医療情報処理エージェントの作成

本研究では以前よりエージェントによる医療情報検索を研究している．しかし研究が発展するとともに，エージェントの機能が情報検索分野以外にも拡大したため，それらを統合する必要があった．そこで複数のエージェントが Web サーバ上で動作している際に，ユーザが Web ブラウザをもちいて指示をあたえられるようにインターフェイスを実装した．本年度は情報検索機能 [4] と文書分類機能の強化もおこなっており，今後は公開できる機能を順次公開し研究に必要なデータを収集する．収集するデータの例として，ユーザから情報検索エージェントに検索キーワードとしてあたえられる一般語があげられる．

4 実験，評価について

本研究で作製した辞書を評価するため，形態素生起コストの学習方法，学習に使用するコーパスなどの条件から辞書を複数パターン作成する．それらの辞書を使用することで得られる形態素解析や文書分類の精度を比較し有効な辞書作成手法を検討する．さらに，有効な辞書作成手法により作成した辞書をもちいて複数の異なる性質をもつコーパスについて分析した結果を，他者が作製した評判の良い形態素解析，文書分類システムによる分析結果と比較し，精度により評価をおこなう．医療情報検索についての評価は既存の情報検索システムではユーザの目的を解決できない非専門家による質問セットを作成し，これをもちいて実験する．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作成した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日本語話し言葉をもちいた医療情報検索の手法として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．今後は実験をおこない作成した辞書の有効性を検証する．課題としては作製する辞書が ipadic を頼る必要がないように充実させることと，形態素生起コストの算出に使用する形態素解析器の発展である．

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] 永田昌明, 「統計科学のフロンティア」第 10 巻 言語と心理の統計, 第 2 部 「確立モデルによる自然言語処理」, 岩波書店, 2003

[3] 北研二, 言語と計算 4 確立的言語モデル, 東京大学出版会, 1999

[4] 北, 津田, 獅子堀, 情報検索アルゴリズム, 共立出版, 2001

[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] 工藤, 山本, 松本, Conditional Random Fields を用いた日本語形態素解析, 情報処理学会研究報告, 2004-NL-161 (13), 2004

[7] 松本, 北内, 山下, 平野, 松田, 高岡, 浅原, 日本語形態素解析システム「茶筌」version2.2.1 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

[8] 浅原, 松本, 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

[9] 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