一般口演 | 知識工学
一般口演10
知識工学
2019年11月23日(土) 09:00 ～ 11:00 C会場 (国際会議場 2階国際会議室)

[3-C-1-02] 医療辞書自動作成システムの構築に向けた医療用語の用語性判定の試み

○櫻井 理紗[1]、竹村 匡正[2]、平松 治彦[1]、上村 幸司[1]、山本 剛[1,2]、奈良崎 大士[1,2]、宍戸 稔聡[1] （1. 国立循環器病研究センター, 2. 兵庫県立大学大学院応用情報科学研究科）

キーワード：natural language processing, text data, web

昨今、医療機関には電子カルテシステムが導入され、医療データが蓄積されるに伴い、臨床研究や大規模データ分析に電子カルテ上に記載されたテキストデータを活用することが期待されている。これらカルテ記載情報を利用するために自然言語処理が行われるが、この自然言語処理については形態素解析器及び「辞書」と呼ばれる用語に対して品詞等が付与される電子ファイルが用いられる。しかし、医療分野等の専門性の高い文章においては、利用される用語の専門性が高くなり、また日々新しい用語が用いられる。その結果、一般的に準備されている辞書では文章の解析精度が担保できず、結果機械学習の適用などの医療テキストデータの二次利用は進んでいるとは言い難い。そのため、新たに出現した用語を自動的に辞書に追加する仕組みが必要である。
一方で、電子カルテのようにテキストデータが電子的に蓄積されることで、用語を自動的に抽出できる可能性がある。例えば、カルテ上でよく記載される単語(文字列)があった場合、これがカルテ上でどれくらい出現しているのか、また他の言語リソース(ウェブやオンラインジャーナルなど)上でも用いられているのか、という知識を利用することで、その単語を単語として抽出すること、すなわち「用語性」の判定を行うことができる可能性がある。
そこで本研究では、カルテ記載情報から、実際のウェブ上の情報を利用して医療用語の用語性の判定が可能かを検証することを目的とする。具体的には、電子カルテ記載に対して形態素解析器を用いて自然言語処理を行い、名詞、連接語および未知語の取得を行う。これら得られた用語をウェブ上の情報に対して APIを活用し問い合せを行うことで、検索ヒット数、スニペット等の詳細情報等の取得を行う。得られた情報から用語性の判定が可能か検証を行うこととする。

医療辞書自動作成システムの構築に向けた医療用語の用語性判定の試み

櫻井 理紗[*1]、竹村 匡正[*2]、平松 治彦[*3]、上村 幸司[*4]、山本 剛[*5]、奈良崎 大士[*6]、宍戸 稔聡[*7]
*1 国立循環器病研究センター、*2 兵庫県立大学大学院応用情報科学研究科

Attempt to determination of termhood for automatic creation of medical terminology dictionary

Risa Sakurai [*1], Tadamasa Takemura[*2], Haruhiko Hiramatsu[*1], Koji Uemura[*1], Tsuyoshi Yamamoto[*1,2], Hiroshi Narazaki[*1,2], Toshiaki Shishido[*2]
*1 National Cerebral and Cardiovascular Center, *2Graduate School of Applied Informatics University of Hyogo

In highly specialized medical texts, the terms used are highly specialized, and new terms are used every day. In particular, highly specialized terms are often compound word, and extracting these and registering them in a dictionary is said to affect the accuracy of applications using natural language processing. On the other hand, by using the knowledge that text data is stored electronically like an electronic medical record, how much it appears on the medical record and whether it is also used on other language resources. Thus, there is a possibility that the word can be extracted as a word, that is, “termhood” can be determined. Therefore, in this study, we extract compound word from medical record data, verify whether it is possible to determine terminology for compound word appearing in the medical record data by using frequency extraction in the medical record data and information on the web.

Keywords: natural language processing, text data, web

1. 背景

昨今、医療機関には電子カルテシステムが導入され、医療データが蓄積されるに伴い、臨床研究や大規模データ分析に電子カルテ上に記載されたテキストデータを活用することが期待されている。これらカルテ記載情報を利用するために自然言語処理が行われるが、この自然言語処理については形態素解析器及び「辞書」と呼ばれる用語に対して品詞等が付与される電子ファイルが用いられる。医療分野の専門性の高い文章においては、利用される用語の専門性が高くなり、また日々新しい用語が用いられる。これらの用語に対して辞書を更新していくことはコストが大きいことから、自動的に辞書更新する仕組みが必要である。特に、専門性の高い用語は複合語であることが多く、これらを抽出し辞書登録することは、自然言語処理を用いたアプリケーションの精度に影響を与えることが言われている[1,2,3]。しかし、電子カルテ記載データに含まれる複合語に対して、それが用語であることを自動的に判定できるかの検証はなされていない。

一方で、電子カルテのようにテキストデータが電子的に蓄積されることで、用語を自動的に抽出できる可能性がある。例えば、カルテ上でよく記載される単語(文字列)があった場合、これがカルテ上でどれくらい出現しているのか、また他の言語リソース(ウェブやオンラインジャーナルなど)上でも用いられているのか、という知識を利用することで、その単語を単語として抽出すること、すなわち「用語性」の判定を行うことができる可能性がある。

2. 目的

そこで本研究では、カルテ（SOAP）記載データから複合語を抽出し、SOAP 記載における頻度、および実際のウェブ上の情報を利用して、複合語の用語性の判定が可能かを検証することを目的とする。

3. 方法

A 病院における一研究で収集された 2 年分の SOAP の記載のうち 16,641 レコードの SOAP を対象とし、形態素解析を行った上で、名詞や接尾辞で構成される複合語の取得を行った。その上で、取得した複合語のカルテ内での出現頻度を算出した。

次に、得られた複合語が、ウェブ上でどのような文脈でどれくらいの頻度で出現しているのか、という情報を利用することによって、用語性の判定に寄与できるのではないかと考え、各用語についてウェブ検索を行うクローラエージェントを構築し、検索ヒット数とスニペットの詳細情報の取得を行った。このスニペットとは、ウェブ検索時に表示される検索語が含まれる文字列のことであり、スニペットを取得することで、その用語がどのような文章で用いられているか把握できる。

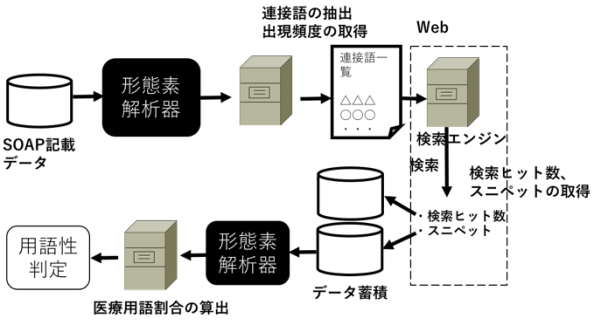

図1 用語性の判定方法

3.1 SOAP 記載における形態素解析および複合語の取得と複合語の出現頻度抽出

SOAP 記載に対して、形態素解析器 MeCab を用いて形態素解析を行った。形態素解析時には、医療用語辞書(英単語 64,936 語、日本語 203,940 語)を使用した。形態素解析で名詞(一般)および接尾辞と判定されたもので構成される複合語を取得し、その出現頻度を導出した。抽出された複合語には、医療辞書に存在する用語のみで構成される用語、医療用語辞書に含まれない用語のみで構成される用語および医療用語辞書に含まれるものとそうでないもので構成される用語が存在する。

また、取得した複合語がカルテ内で出現する頻度を算出した。

3.2 Web 上のデータ収集のためのクローラーの構築

取得した複合語を一語ずつ自動的に検索エンジンに問い合わせ、検索ヒット数とスニペットを取得するクローラエージェントによるデータ収集システムを構築した。本システムは Python 言語を使用した。具体的には、検索した結果のページの HTML の記述を取得し、その取得したデータの中から、検索結果とスニペットを抽出し、データの蓄積を行った。今回、検索エンジンにはマイクロソフト社が提供する Bing を使用した。

3.3 スニペット内の医療用語の割合

各複合語に対して得られたスニペット内の文章を形態素解析を行い、スニペット内における医療用語の割合を算出した。医療用語判定には先述した医療用語辞書によって判定した。

4. 結果

SOAP 記載から取得した複合語は 8,499 語であり、SOAP 内記載内の総複合語数は 42,736 語であった。最も高頻度に出現した単語は、「行動体」などの2つの複合語が 737 回出現していた。逆に、「高負荷増大」を始めとする 5069 個の複合語は一回出現下のみであった。全体としては、出現頻度 1 回のものが全体の 60%を占めていた。SOAP 記載内の出現頻度を表したものを図 2 に示す。

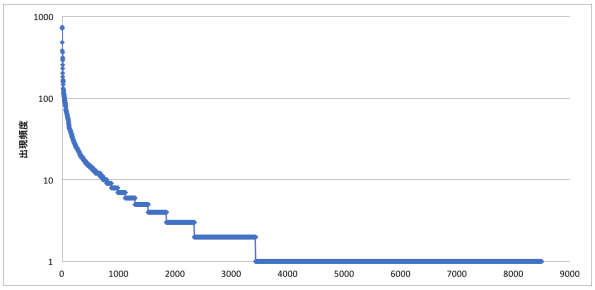

図 2 SOAP 記載内の出現頻度

4.1 高頻度語の用語性判定

出現した複合語のうち、任意の 10 回以上出現した複合語 100 個に対して、人手で医療用語といえるかどうか、すなわち用語性の判定を行った。その結果、100 語のうち、87 語(87%)が用語性ありと判定された。

4.2 低頻度語の用語性判定

次に SOAP 記載における出現回数が 10 回未満の複合語に対して、用語性の判定を行った。具体的には、高頻度語と同じく任意の 100 個の複合語を抽出し、目視で確認を行った。その結果、69 語(69%)が用語性あり、と判定された。

4.3 ウェブ検索エンジン上での出現分析

次に、抽出された複合語に用語性があるかどうかについて、検索エンジン上での検索ヒット数、および医療用語の出現割合について検討した。検索ヒット数については、ばらつきが大きいため対数をとり、ヒストグラムを作成したところ、以下のようになった。

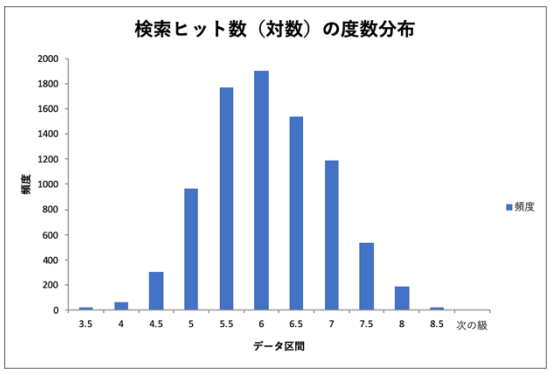

図 3 検索ヒット数(対数)の度数分布

「用語性あり」とみなされた複合語の検索ヒット数と、「用語性なし」とみなされる複合語の検索ヒット数(の対数)について、t 検定を行ったところ、有意差は認めなかった。また、医療用語の出現割合についてもウィルコクソンの順位和検定を行ったところ、こちらも有意差を認めなかった。

5. 考察

SOAP 記載内に出現する複合語は、今回医療用語辞書を用いて形態素解析を行ったこともあり、多くが医療用語として認められた。特に、高頻度で出現する複合語は、高い確率で医療用語として認められることがわかった。一方で、低頻度で出現する複合語についても、高頻度複合語までではないにしても、多くの医療用語が含まれていることがわかった。これらの峻別については、ウェブ上での利用頻度、医療用語として用いられているかの情報を利用すれば、より高い精度で峻別できると予想したが、明らかな有意差は認めなかった。一方で、用語性の判定には、SOAP 上の用法に合わせてウェブ上での用例を確認することが有益であることから、候補となる複合語の用例を判別する新たな方法を検討する必要があると考えられる。

参考文献

1) 中川裕志，湯本紘彰，森辰則．出現頻度と連接頻度に基づく専門用語抽出．自然言語処理 2003 ; 10(1) :27-46.
2) 青木和夫，中山章弘，松崎剛士．形態素解析での効率的な複合語処理．情報処理学会研究報告．1-6, 2003.
3) 小山照夫，竹内孔一，専門用語抽出における形態素辞書変更の効果，情報処理学会研究報告，1-4, 2014