1F4-04

病状経過知識抽出のためのテキストマイニング

Mining a Clinical Cource Knowledge from Summary Text

阿部 秀尚[*1] 平野 章二[*1] 津本 周作[*1]
Hidenao Abe Shoji Hirano Shusaku Tsumoto

[*1]島根大学医学部医学科医療情報学講座
Department of Medical Informatics, Shimane University School of Medicine

In this paper, we present a text mining method to extract clinical cource knowledge from documents, which are written in natural langguage. To identify medical terms in these documents, we have done morphological analysis with MEID dictionary. Then we have applied association rule learning method to extract association rules, which present some clinical cource about neuro-phisical diseases. With this result, we disscuss about tecnical issues to extract clinical knowledge from medical documents.

1. はじめに

近年，医療現場における検査結果や診療情報の電子化が進み，例えば，1 日に 1000 人規模の外来をもつ大学附属病院においては 1 年にテキストデータとして 20GB 以上のデータが蓄積され，その活用が課題となっている．また，個々の診療における質を向上させるため，科学的根拠に基づく診療 (EBM:Evidence Based Medicine) が注目され，ここでの科学的根拠の多くは PubMed などによる文書の検索によって得られているのが現状である．しかし，これらの情報検索ではキーワードによる検索が一般的であり，目的の知識を得るには複数の文書を読み比べるなど，人的コストが高い．このような現状を改善するため，医学文書における知的な検索の支援手法の開発が望まれている．

以上の課題に対し，本研究では，テキストマイニング技術を用いて病状経過や治療効果に関する知識を語彙間の関係として抽出する試みを行った．本稿では，島根大学医学部附属病院から提供された退院時サマリーを用いて，病状経過に関する特徴的な語彙間の関係を抽出し，結果と今後の課題について考察する．

2. 退院時サマリーからの病状経過知識抽出

本研究では，自然文で記述された病歴情報から，病状経過・治療効果に関する知識の抽出を行うため，自然言語で記述された文書に対してテキストマイニング手法を適用し，特徴的な語彙間の関係を抽出する．ここでは，退院時サマリーと呼ばれる文書を用いて，語彙の同定を行い，語彙間の関係を相関ルール [Agrawal 94] によって抽出する．

2.1 退院時サマリーの概要

本実験で用いる退院時サマリーは，患者の退院時に入院中の状況を簡潔にまとめた情報であり，疾患や経過に関する情報を表現している．ここでは，島根大学医学部附属病院から提供された神経内科に関する疾患を主訴とする 10 症例の退院時サマリーを対象とした．これらの退院時サマリーは表 1 に示すような項目から構成されている．表 1 中の「不完全な文」は，2 語以上の自立語や記号を含むが，助詞などの付属語を含まない文であることを意味する．「完全な文」は，それ以外の文を意味し，主語や述語が整っていない文を含む．

表 1: 退院時サマリー内の項目と内容

項目	内容	
	不完全な文	完全な文
診断	○	—
主訴	○	—
既往歴	○	—
家族歴	○	—
現病歴・現症	○	○
入院時検査所見	○	○
入院後の経過	○	○
退院時処方	○	—

ここで,「診断」「主訴」は疾病の名称だけから成り,「既往歴」は時期と疾病の名称,「家族歴」は本人との関係と疾病の名称が記入されている.「現病歴」は主な症状についてそれまでの病歴をまとめた文章である．これには，加療の履歴などが含まれる.「入院時検査所見」は入院時に行う血液・尿・画像の検査所見が検査項目と結果の対，あるいは自然言語により記入されている.「入院後の経過」は入院後の治療の履歴と病状の経過が箇条書きまたは文章で記述されている.「退院時処方」は処方した薬剤が列挙されている．

2.2 前処理

日本語で記述された文書を扱う場合，はじめに形態素解析により，単語の同定と各語の品詞の同定を行う必要がある．本研究では，形態素解析ソフトウェア「茶筌」[ChaSen] を利用した．

まず，茶筌によって，各文の形態素解析を行う．茶筌の付属辞書には新聞記事などに用いられる一般的な単語が登録されているが，専門性の高い医学用語の識別は難しい．このため，茶筌による形態素解析の結果だけでは，医学用語を含む文の解析は十分とは言えない．また，茶筌では単語の解析について，動的計画法と学習による形態素解析が用いられており，各単語には遷移のしやすさを登録するため，正しい形態素解析結果が大量に必要である．このため，医学用語辞書を単純に茶筌に登録するだけでは，正しい形態素解析が行える保証がない．

本実験では，医学用語辞書 [MEID] に問い合わせを行い，医

連絡先: 阿部 秀尚，島根大学医学部医学科医療情報学講座，〒693-8501 島根県出雲市塩冶町 89-1，0853-20-2174，0853-20-2170，abe@med.shimane-u.ac.jp

学用語を同定する．茶筌による形態素解析の結果では，医学用語が名詞句として得られるため，これらを結合し医学用語辞書の索引であるかどうかを問い合わせる．ただし,「高血圧」と「血圧」のように 2 つの医学用語が存在することもあるため，単語の長さを優先した競合解消を行った．

なお，日本語によって記述された文書における医学用語を正しく認識するためには，ここであげた経験則的な方法ではなく，文書中の単語頻度などを用いて認識する手法 [竹内 04] など議論の余地がある．

2.3 相関ルールからの語彙関係の抽出

2.2 で述べた医学用語の識別を 10 症例について行った結果を用い,「入院後の経過」についてデータセットを構築した．この項目に含まれる各文をインスタンスとし，全文に含まれる語彙を属性とする．データセットの概要を表 2 に示す.

表 2: データセットの概要

症例 No.	退院時	インスタンス数	語彙数
1	不変	5	14
2	不変	15	55
3	その他	10	44
4	その他	25	43
5	その他	12	47
6	軽快	17	55
7	その他	60	99
8	不変	12	50
9	軽快	6	23
10	軽快	7	31

このデータセットに対し，相関ルール学習 [Agrawal 94] を適用し，得られた相関ルール $X \Rightarrow Y$ について，$P(X|Y)/P(Y|X) > 1$ となるルール対を取りだす．相関ルールの実行には Weka[Witten 00] を用い，各データセット D_i に対して最小支持度として $2/|D_i|$ を，最小確信度として 0.5 を与えた.

症例6：

IF CT THEN **腹部大動脈瘤** (support(2/17), confidence(1.0))

IF **腹部大動脈瘤** THEN CT (support(2/17), confidence(0.67))

症例7：

IF **下痢** THEN WBC (support(2/60), confidence(1.0))

IF WBC THEN **下痢** (support(2/60), conficence(0.6))

図 1: 各症例から得られた語彙間の関係を示す相関ルール

図 1 は，以上の結果，得られたルールを示す．この結果，症例 6 からは,「CT(Computer Tomography)」という検査によって「腹部大動脈瘤」という疾患を検査するというルールが得られた．症例 7 からは「下痢」という症状について「WBC(白血球数)」という検査項目に着目していたことを示すルールが得られた．これらのルールは，医師が行った行為を表しているものと考えられ，それぞれの症状についてどのような行動をとれば良いのかを明示している．

3. 考察

以上の実験より得られた退院時サマリーからの病状知識抽出について考察する．

まず，前処理に関して，今回は形態素解析の結果のみを用いて医学用語の同定を行った．しかし，医療に関する文書では，検査項目と検査結果の対，ある事象の観測結果などが記述されていることが多く，これらを積極的に用いる必要がある．

次に，語彙間の関係について，今回は個々の症例について記述された文単位に存在する語彙間の関係を抽出しようと試みた．しかし，より多くの文書から一般的な語彙間の関係を抽出し，疾患に特有の語彙の組み合わせを抽出する方法などが考えられる．疾患に特有の語彙抽出については，$tf \times idf$ を用いた手法 [小野 04] などがあるが，これをさらに進めて語彙間の関係抽出を行うことも考えられる．

最後に得られた語彙間の関係の利用については，医学教科書 (メルクマニュアル) などから語彙間の関係を抽出し，これを標準的な知識として蓄積することが考えられる．また，標準的な知識に対して，自然言語による要求を可能にするため，退院時サマリーのような短い文書内に存在する語彙間の関係を抽出する手法の開発を進める必要がある．

4. おわりに

本稿では，自然言語によって記述された医療文書である退院時サマリーからテキストマイニングによって病状経過や治療効果についての知識の抽出を試みた．相関ルールによる医学用語集合からの語彙間の関係抽出実験では，いくつかの特徴的な語彙間の関係を抽出した．

今後は，自然言語処理の技術を用いた医学用語だけではなく，検査結果や治療行為を含めた語彙間の関係から病状経過に関する知識の抽出を行う必要がある．また，さらに大規模に文書を収集し，テキストマイニングによる病状経過知識の抽出を行っていく．

参考文献

[Agrawal 94] Agrawal, R., and Srikant, R.: Fast algorithms for mining association rules in large databases, In Proc. of International Conference on Very Large Data Bases, pp. 478-499 (1994).

[ChaSen] 茶筌：http://chasen.org/

[MEID] 25 万語医学用語大辞典，日外アソシエーツ，(2005).

[Witten 00] Witten, I. H and Frank, E.: DataMining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, (2000).

[小野 04] 小野 大樹，高林 克日己，鈴木 隆弘，横井 英人，井宮 淳，里村 洋一:テキストマイニングによる退院サマリー自動分類の試み，医療情報学，24(1), pp. 35-44 (2004).

[竹内 04] 竹内 匡正，松井 弘子，芦田 信之:用例に基づく医療用語知識の体系化について，医療情報学，24(1)，pp. 139-145 (2004).