3B3-NFC-01b-1

テキストデータマイニング統合環境を利用した看護記録からの専門用語辞書作成支援ツールの提案

Proposal on Technical Term Dictionary Creation Support Tool Using TETDM

高間 康史 *1　　阿部 美里 *2
Yasufumi Takama　　Misato Abe

*1首都大学東京大学院システムデザイン研究科
Graduate School of System Design, Tokyo Metropolitan University

*2首都大学東京システムデザイン学部
Faculty of System Design, Tokyo Metropolitan University

This paper proposes a support tool for creating technical term dictionary from nursing reports. A technical term dictionary is one of important components of text mining system, which affects overall performance of the system. Therefore, it is important to support dictionary creation. The proposed tool is implemented using TETDM (Total Environment for Text Data Mining), which is suitable for user-participated tool development. In order to make use of such a characteristic, the proposed tool consists of a module that is designed to be easily replaced with other modules, and those with general-purpose functionalities. This paper explains functionality of each module as well as coordination among those. The result of extracting technical terms from actual nursing reports using the proposed tool is also shown.

1. はじめに

本稿では，テキストデータマイニング環境 TETDM (Total Environment for Text Data Mining)[砂山 13] を利用し，看護記録から専門用語候補を抽出して専門用語辞書を作成する作業を支援するツールを提案する．看護記録は患者の状態と共に，看護行為の目的や必要性の判断，実施内容などを記載したものであり，医療・看護の継続性や，医療従事者と患者間での診療情報共有に繋がる重要な文書である [厚労省 05]．近年，看護記録へテキストマイニングを適用し，質的監査を効率的に行うことなどが期待されているが，現状では看護記録へのテキストマイニング事例は少なく，まだ試行の段階といえる [村松 10]．

本稿では，テキストマイニングの性能を左右する重要な要素の一つである専門用語辞書の構築を対象とする．実際の看護記録から候補語を自動抽出し，人手で確認しながら辞書を構築する作業を支援する．提案ツールの開発に採用する TETDM は，複数のモジュールを組み合わせ，一つのツールとして動作させることが可能であり，ユーザ参加型のツール開発に適している．この特徴を活かすため，本稿では拡張性を考慮したモジュール構成を検討する．具体的には，専門用語抽出，抽出単語の提示，看護記録表示，看護記録中の用語ハイライトの 4 つの機能に分割し，それぞれモジュールとして構築する．実際に構築したツールのプロトタイプを用いて用語抽出を行った結果について示すと共に，拡張可能性について考察する．

2. TETDM

TETDM[砂山 13] は，テキストデータマイニング及びその処理結果の可視化ツールの開発を容易にすること，および開発されたモジュールを広く公開し，多数ユーザに利用してもらうことを目的として提案され，現在も開発が続けられている．図 1 に TETDM のスクリーンショットを示す．統合環境は複数のパネルで構成されており，各パネルはテキストデータマイニング処理を行う**マイニング処理モジュール**とその出力を可視化する**可視化インタフェースモジュール**の対から構成される．パネルに割り当てるマイニング処理モジュール，可視化インタフェースモジュールはそれぞれ，リアルタイムに切り替えることができる．

図 1: 専門用語辞書作成支援ツールのスクリーンショット

マイニング処理モジュールと可視化インタフェースモジュールは単体では機能せず，それぞれパートナーとなるモジュールを必要とする．しかし，必ずしも 1 対 1 で開発される必要はなく，データのやりとりの仕様が共通化されていれば，一つの可視化インタフェースモジュールが複数のマイニング処理モジュールに対応することや，その反対も可能である．この特性を生かし，可視化インタフェースモジュールは共通のまま，異なるマイニング処理モジュールに切り替えることで，同一インタフェースで多様なデータ分析作業の支援が可能となる．また，異なるアルゴリズムの比較実験を行う場合にも有効と考える．反対に，同一のマイニング処理モジュールに対し異なる可視化インタフェースモジュールを実装することで，インタフェースの比較環境としても有効活用できることが期待できる．

3. 専門用語辞書作成支援ツール

提案する専門用語辞書作成支援ツールのモジュール構成を図 2 に示す．ツールは二つのパネルから構成され，一つはマイニング処理モジュール TermExtraction と可視化インタフェース

連絡先: 高間 康史，首都大学東京大学院システムデザイン研究科，〒 191-0065 東京都日野市旭が丘 6-6，Tel/Fax. 042-585-8629，ytakama@sd.tmu.ac.jp

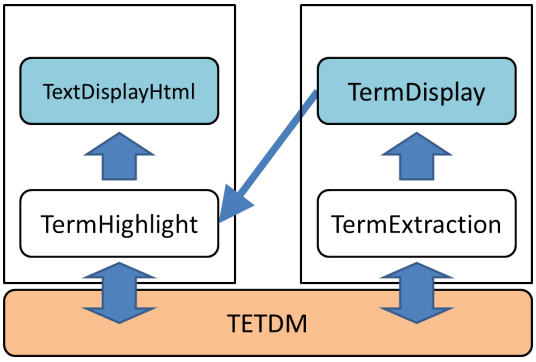

図 2: 専門用語辞書作成支援ツールのモジュール構成

モジュール TermDisplay，他方はマイニング処理モジュール TermHighlight と可視化インタフェースモジュール TextDisplayHtml からそれぞれ構成される．図 1 は TETDM を用いて実装したプロトタイプのスクリーンショットであり，左側が TextDisplayHtml，右側が TermDisplay のパネルにそれぞれ対応する．

TermExtraction モジュールは，後述する C-Value などの専門用語抽出手法を実装し，スコアの降順にソートした専門用語候補のリストを TermDisplay モジュールに送る．ツール利用者は，TermDisplay モジュール上に表示された候補語を見ながら，それが専門用語として適切か否かを選択できる．候補語上でダブルクリックするたびに，状態が正解語・不正解語・未チェックの順に変化するようになっており，Output ボタンを押すことによって，正解語とそのスコアをテキストファイルに出力する．

TermHighlight モジュールは，TermDisplay モジュール上でシングルクリックされた候補語の看護記録中での出現箇所をハイライトし，その結果を HTML ソースとして TextDisplayHtml モジュールに送る．TextDisplayHtml モジュールは HTML ソースを表示するモジュールであり，TETDM に標準で付属する．TermDisplay と TermHighlight 間の連携は，TETDM のフォーカス連動機能を利用して実装している．

本稿で実装した専門用語抽出手法は C-Value[中川 03] であり，構成単名詞数，出現回数等に基づき複合名詞を評価する．単名詞の場合にスコアが 0 にならないように C-Value を修正した MC-Value も提案されている [湯本 01]．TETDM で形態素解析を行い，キーワードに分類されたものを単名詞として単語 N-Gram により候補語リストを作成し，C-Value を求める．

TermExtraction モジュールは，各種専門用語抽出アルゴリズムを実装し，差し替えて利用可能とするため，ベースとなるモジュール（クラス）として TermMiningBase を開発した．このモジュールは TermDisplay との連動に関する処理のみを実装したものであり，マイニングモジュールはこれを継承することで開発する．

4. 動作検証

実際の看護記録 1，5，10，20，50，100，200 件をそれぞれ TETDM の入力として，構築したツールの動作検証を行った．看護記録の件数と候補語数，うちスコアが 0 でない語数の関係をグラフにしたものを図 3 に示す．また，看護記録 5，20，200 件を入力した際のスコア上位の 5 語を表 1 に示す．実行速度に関しては，Intel(R) Core(TM) i7-2600，メモリ 8GB，Windows7 64bit，JDK1.7.0_10（64bit 版）で，看護記録 50 件までは 5 秒以下，100 件で 50 秒，200 件で 210 秒であった．

表 1: スコア上位 5 語の比較

順位	看護記録 5 件	看護記録 20 件	看護記録 200 件
1	潰瘍性大腸炎	潰瘍性大腸炎	内視鏡
2	内視鏡	内視鏡	自己免疫性溶血性貧血
3	性大腸炎	潰瘍性大腸	免疫性溶血性貧血
4	潰瘍性大腸	性大腸炎	自己免疫性溶血性
5	消化管内視鏡検査 潰瘍性大腸炎	炎症性腸疾患	潰瘍性大腸炎

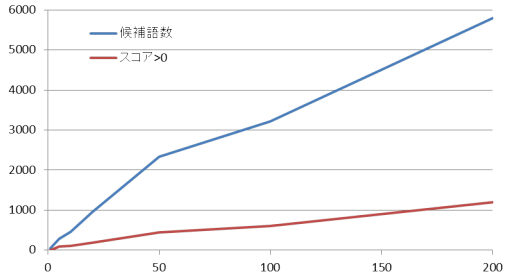

図 3: 看護記録件数と候補語数の関係

前述の通り，TermExtraction モジュールを差し替えることで，ユーザは同一のインタフェースのまま他の用語抽出アルゴリズムを用いることが可能である．また，TETDM は任意枚数のパネルを同時に表示できるため，各 TermExtraction モジュールのパネルを同時に表示し，比較しながら作業することも可能である．さらに，TermDisplay モジュールを差し替えることによるインタフェースの変更や，副次的情報を表示する他のパネルを追加するなどの拡張も可能である．

5. おわりに

本稿では，TETDM を用いた専門用語辞書作成支援ツールを提案し，実際の看護記録に適用して動作検証を行った．現在，C-Value 以外の代表的な用語抽出手法についてもモジュールを実装しており，今後公開する予定である．

参考文献

[砂山 13] 砂山，高間，西原，徳永，串間，阿部，梶並，テキストデータマイニングのための統合環境 TETDM の開発，人工知能学会論文誌，Vol. 28, No. 1, pp. 1-12, 2013.

[厚労省 05] 厚生労働省，検討会における「看護記録について」に関して出された主な意見，http://www.mhlw.go.jp/shingi/2005/10/s1017-12.html （2013 年 4 月 4 日現在）．

[村松 10] 村松，渡部，大崎，小塚，看護記録のテキストマイニング，情報処理学会論文誌，Vol.3, No.3, pp. 112-122, 2010.

[中川 03] 中川，森，湯本，出現頻度と連接頻度に基づく専門用語抽出，Journal of natural language processing, Vol. 10, No. 1, pp. 27-45, 2003.

[湯本 01] 湯本，大畑，森，中川，語基の連接情報を用いた専門語抽出，言語処理学会年次大会発表論文集，Vol. 7, pp. 161-164, 2001.