1L−4

医療専門用語登録支援システムのための専門用語抽出・分類手法

坂田 大輔　永井 秀利　中村 貞吾　野村 浩郷　大貝 晴俊†　中島 律子‡

九州工業大学大学院 情報工学研究科　新日本製鉄† 科学技術振興団‡

1 はじめに

我々は医療テキストからの情報抽出に関する研究[1]を行なっている．自然言語の計算機処理において，新語や未知語の出現は情報抽出において処理を複雑にしている．そのためテキスト中から新語や未知語を抽出し，それらの語に専門用語の属性を付与し計算機辞書に登録するシステムが有用である．

専門用語の属性はその用語の語構成に大きな手掛かりがある．また，専門用語の属性は意味や用途に応じて設定されているため，文章中に出現する際，専門用語は他の同じ属性の用語と似た文法で出現したり，共通の語が共起したりと，各属性ごとに特徴がある．

そこで今回，専門用語の語構成，係受け情報を用いて，テキストから未知語，複合語を抽出し，その語の属性を特定する手法を提案する．そして，それらの抽出分類手法を実装した医療専門用語登録支援システムを作成した．

2 医療論文抄録と分類属性

対象テキストとして日本医学放射線学会の学術発表会抄録を用いる（1記事 平均 12.0 行 736.7 文字）．今回は表1に示される専門用語の細分類属性に準じて分類を行なう．

表 1: 専門用語細分類属性

細分類属性	例
病名	Ｃｒｏｈｎ病，急性骨髄単球性白血病
性質	解離性，偽腔開存型
診断方法	血流予備能計測，ＭＲＩ形態診断
診断ソース	３Ｄ画像，ＭＲＩ所見
診断単位	症例，ｍｍｏｌ／ｋｇ
診断尺度	腫瘍径，スライス厚
診断機器	液体加温注入装置，内視鏡
治療方法	経静脈性塞栓術，マイクロ波凝固療法
解析方法	２Ｄ撮像法，ＲＯＣ解析
診断対象	肝細胞癌患者，右気管支動脈
現象	疼痛改善，形態変化
医療物質	ホルマリン液，ジエチルエーテル麻酔

3 システム概要

3.1 医療専門用語の抽出分類

以下の手順で専門用語の抽出分類を行なう．

1. 分類対象語の抽出
   まず，抄録テキスト中から対象となる未知語，複合語を抽出する．

2. 語構成情報による分類
   専門用語の中から属性ごとに頻出する語尾を抽出し，これらを語尾パターン（表2）とする．この語尾パターンを分類対象語とパターンマッチを行ない，一致する場合，その語は語尾パターンの属する属性に分類される．

表 2: 専門用語の属性別語尾パターン

細分類属性	例
病名	癌，腫瘍，疾患
性質	性，型，的
診断方法	検査，診断，測定
診断ソース	像，所見，影
診断尺度	率，効果，期
診断機器	ＣＴ，装置，システム
治療方法	療法，術
解析方法	法，式，分類
診断対象	脈，細胞，管
現象	化，集積，停止
医療物質	線，因子，液

診断単位の抽出に関しては連接パターンを用いることにより抽出を行なった．また“Siemens Magnetom Vision 1.5T”というように，中心となる機器名の後に“1.5T”といった機種名等が語尾につくものも見られたため，語尾パターンを分類対象語とパターンマッチを行い，マッチした部分が語尾でない場合はその語の語尾が連接可能であるかを特定し抽出した．

また，3文字のアルファベット列においては病名，診断機器，診断方法，解析方法等のいずれかであることが多いことや，4文字以上のアルファベット列は単位となることが稀である等といった，属性を限定することのできる傾向が見られる．この傾向を利用して分析抄録中のアルファベット列の文字列長別の分類属性別の出現確率をスコアリングの基準の一つとして用いた．

3. 係受け情報による分類
   分類対象語の係り先となる文節，分類対象語に係る文節によりスコアルールを作成する．分析抄録から専門用語を抜き出し，その専門用語の係り先となる文節，専門用語に係る文節を抜き出す．それぞれの抜き出された文節に対して，文節と専門用語の属性別の共起確率を求める．それらの文節と属性別の共起確率の組みを係り受けスコアルールとする．

   分類対象語に対してこれらのスコアルールが複数存在する場合はこれらのスコアを掛け合わせることで分類対象語の属性スコアを導き出す．そして最も高いスコアを持つ属性を分類対象語の細分類属性とする．

3.2 医療専門用語登録支援システム

3.1で述べた分類手法を実装した医療専門用語登録支援システムを作成した（図1）．これは抄録テキスト

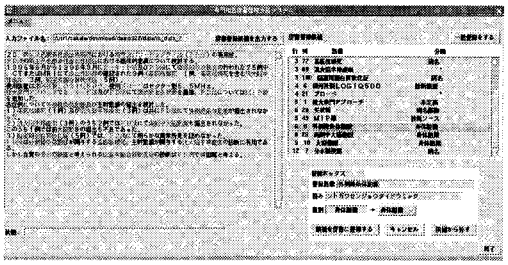


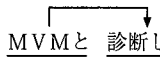
図 1: 登録支援システム実行画面

ファイルを入力することにより，テキスト中の専門用語の候補語をアピール・一覧表示したり，それらの語を選択し専門用語として編集，登録することができる．さらに登録した専門用語の結果を再び登録システムに反映させることができる．

4 結果

4.1 分類例

以下の文における未知語 “ＭＶＭ” の分類例を示す．



ＭＶＭと 診断した ３７患者の ３８病変で，

“ＭＶＭ” は “と” を伴って，“診断した” に係り，どの文節からも係り受けることのない，３文字の大文字アルファベット文字列である．ここでは，まず “ＭＶＭ” は “ ～ と 診断した” という係り受けスコアルールにマッチする．“ ～ と 診断した” の各属性の出現確率はスコアリングデータより

$$[p_{\text{病名}}, p_{\text{性質}}, p_{\text{方法}}, p_{\text{ソース}}, \cdots\cdots, p_{\text{物質}}]$$
$$= [\ 0.67, 0.03, 0.05, 0.02, \cdots\cdots, 0.00]$$

と与えられている．また，“ＭＶＭ” は “アルファベット (大文字) 列３文字” のスコアルールにマッチする．“アルファベット (大文字) 列３文字” の各属性の出現確率はスコアリングデータより

$$[p_{\text{病名}}, p_{\text{性質}}, p_{\text{方法}}, p_{\text{ソース}}, \cdots\cdots, p_{\text{物質}}]$$
$$= [\ 0.22, 0.00, 0.01, 0.01, \cdots\cdots, 0.13\ ]$$

と与えられている．よって，このときの “ＭＶＭ” のスコアＳは

$$\begin{aligned} S &= [p_{\text{病名}}, p_{\text{性質}}, p_{\text{方法}}, p_{\text{ソース}}, \cdots\cdots, p_{\text{物質}}] \\ &= [0.67 \times 0.22, 0.03 \times 0.00, 0.05 \times 0.01, \\ &\qquad 0.02 \times 0.01, \cdots\cdots, 0.00 \times 0.13\ ] \\ &= [0.1474, 0.00, 0.0005, 0.0002, \cdots\cdots, 0.00\ ] \end{aligned}$$

となり，Ｓの要素中の最大値は $s_{\text{病名}}$ の 0.1474 なので “ＭＶＭ” の属性は病名と判定した．

4.2 医療専門用語の抽出分類

3.1 で述べた手法を用いて専門用語抽出分類実験を行なった．日本放射線学会の論文抄録のうちスコアリングデータ作成 (分析) 用に 495 稿，検証用に 54 稿の抄録を用いた．抽出分類の再現率，適合率は表 3 のようになった．

表 3: 専門用語の抽出分類実験

細分類属性	再現率 [%]	適合率 [%]
病名	78.9	81.9
性質	35.2	69.0
診断方法	53.6	91.6
診断ソース	73.2	65.9
診断単位	42.8	27.3
診断尺度	69.5	75.2
診断機器	74.7	41.7
治療方法	55.4	78.3
解析方法	59.8	76.9
診断対象	76.0	79.0
現象	48.8	72.7
医療物質	37.2	78.8
合計	64.3	69.0

4.3 医療専門用語登録支援システム

今回作成した医療専門用語登録支援システムを用いて専門用語の登録が行なわれ，既存の辞書 (7705 語) に加え，851 語の専門用語が本システムにより登録された．本システムの使用にあたって，試用者からは以下のような意見が挙げられた．

- どの位の専門用語が登録されているか，ひとつの論文の登録状況から把握できた．
- 使いやすく，計算機が使用可能な形で出力されるため，作成時間の短縮につながった．

5 考察

分類属性の診断機器においては，抄録の著者により語の前後に診断機器の詳細な種類を記述するなどといった表記の揺らぎが多く見られた．このような表記の揺れが生じている語を専門用語として登録するのは適当であるか判断し難いが，これらが専門用語であるということを登録支援システムにおいて指示する必要はある．

今回は専門用語の語構成，係受け情報を用いてテキストからの専門用語の抽出と分類を行う支援システムにより，短時間で有効に専門用語辞書を作成させることが可能となった．専門用語の抽出分類については，類似する分類属性 (診断方法と解析方法) では解析方法と分類されなければならないところを診断方法と誤って分類される例が見られた．これらの２つの属性において表層的な語構成や係受けで共起する語が類似する傾向がみられた．人手においても一文読んだだけでは，どちらに分類するべきか判定することは難しく，前後数行の読んで分類できるという場合もいくつか見られた．そのため類似性の高い分類属性に関しては，文脈中の主題を捉えることが必要であると考えられる．

参考文献

[1] 大貝晴俊 他，医療情報の収集とファクトデータの抽出，情報メディア学会第２回研究会，2001

謝辞

この研究は科学技術振興団「高度医療ネットワークに関する研究」の支援を受けて行なわれました．ここに深く感謝の意を表します．