原著

Support Vector Machine（SVM）を用いた自然文読影レポートからの医学的知識の抽出

安永　晋[*1]　川上洋一[*1]　笹井浩介[*1]

我々は，過去のレポートから医学的知識を抽出し，これを用いて入力支援情報を提示することが可能なレポーティングシステムの開発に取り組んでいる．このシステムでは，まず既存の自然文読影レポートを自然言語処理で構造化し，次に，部位・モダリティに応じて，構造化したデータの統計情報を利用して候補単語を入力支援情報として提示する．そして，ユーザーは提示された候補単語を選択することによって読影レポートを作成する．

このシステムに用いる構造化処理として，形態素解析された自然文に「Support Vector Machine（SVM）」と呼ばれるアルゴリズムを適用し，単語に属性を付与するという手法による高精度な構造化処理の実現可能性について検討を行った．

本論文では，まず過去のレポートからこの「SVM」を用いて「部位」「所見」などの「医学的知識」を抽出する技術の詳細と実験結果について述べ，次に「医学的知識」の抽出精度を高めるために「言い換え」を用いる手法の提案とその実験結果について述べる．

■キーワード：Support Vector Machine（SVM），自然言語処理，言い換え，読影レポート

Extraction of Medical Knowledge That Used Support Vector Machine（SVM）from Diagnostic Report of Natural Sentence：Anei S，Kawakami Y，Sasai K

We are working on the development of the reporting system that can present input support information by extracting medical knowledge from a past report，and using this.

In this system，first of all，the existing，natural sentence interpretation of radiogram report is structurized because of the natural language processing. Next，the candidate word is presented as input support information according to the part and the modality by using the statistical information of the structurized data. And，the user makes the interpretation of radiogram report by selecting the presented candidate word.

As structurizing processing used for this system，We examined the realizability of highly accurate structurizing processing by the technique of applying the algorithm that was called “Support Vector Machine（SVM）” in the natural sentence resolved to the morpheme，and giving the word the attribute.

In this thesis，details of the technology that extracts “Medical knowledge” such as “Part” and “Opinion” by first using this “SVM” from a past report and the experiment result of it are described. Next，the proposal of the technique that uses “Paraphrase” to improve the extraction accuracy of “Medical knowledge” and the experiment result of it are described.

Key words：Support Vector Machine（SVM），Natural language processing，Paraphrasing，Diagnostic report

[*1] コニカミノルタテクノロジーセンター株式会社
イメージシステム開発室
〒569-8503　高槻市桜町 1-2
E-mail：anei@eie.konicaminolta.jp
受付日：平成 18 年 2 月 8 日

[*1]Imaging System R&D Division，Konica Minolta Technology Center，Inc.
1-2 Sakura-machi，Takatsuki-shi，Osaka，569-8503，Japan

1．はじめに

近年，病院内の情報システム化が進んで大量のデータを蓄積することが可能となりつつある．しかし，現在の情報システムではサブシステム間におけるデータの相互関係が定義されていないので，相互に関連する知識を効率的に抽出することができない．

そこで我々は，様々なデータソースからデータエレメントを抽出し，「RDF（Resource Description Framework）」を用いて各データエレメント間の相互関係を定義することにより，ユーザーに有用な情報を提供するシステムの開発を行っている[1]．その一例として，過去のレポートから医学的知識を抽出し，これを用いて入力支援情報を提示することが可能なレポーティングシステムの開発に現在取り組んでいる．レポートは基本的に自然文なので，そのままでは医学的知識を抽出するのは困難である．そこで，自然言語処理を用いて文を構造化する必要がある．

我々が開発しているレポーティングシステムでは，まず症例データベースに蓄えられている自然文読影レポートを自然言語処理で**図 1** に示すように構造化し，次に，構造化したデータの出現頻度，共起確率といった統計情報を考慮して候補単語を入力支援情報として提示する．そして，ユーザーは提示された候補単語を選択することによって読影レポートを作成する[2]．

自然文読影レポートを構造化する手法としては，あらかじめ「部位」「所見」などの「医学的知識」を表す単語をすべて記憶した「医学的知識データベース」を作成しておき，その上で自然文を形態素解析して単語単位に分割し，出現した単語の中で「医学的知識データベース」に含まれているものを抽出するという単純な方法も考えられる[3]．しかし，この方法だと「医学的知識データベース」に漏れがあった場合は漏れている単語を抽出することができず，また別の問題として，同じ単語が医学的知識と医学とは無関係な事柄の 2 種類（あるいはそれ以上）の意味を持つ場合に，医学とは無関係な事柄を表す場合でもその単語を抽出してしまうという問題もある．

そこで，形態素解析された自然文に対して，「Support Vector Machine（SVM）」と呼ばれるアルゴリズムを適用し，自然文中の単語に属性（「部位」「所見」など）を付与するという手法による高精度な構造化処理の実現可能性について検討を行った．

本論文では，まず過去のレポートからこの「SVM」を用いて「部位」「所見」などの「医学的知識」を抽出する技術の詳細と実験結果について述べ，次に「医学的知識」の抽出精度を高めるために「言い換え」を用いる手法の提案とその実験結果について述べる．

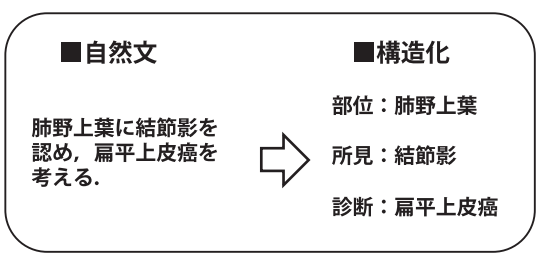

図 1　自然文読影レポートの構造化の例

2．Support Vector Machine（SVM）を用いた医学的知識の抽出

1）Support Vector Machine（SVM）について

「Support Vector Machine」（以下，SVM と記す）とは，簡潔に言えば，過去の事例の学習によってベクトルを二値あるいは多値に分類する方法である[4]．以下に二値の場合の例を示す．

m 個の n 次元事例ベクトル X1, . . . , Xm のそ

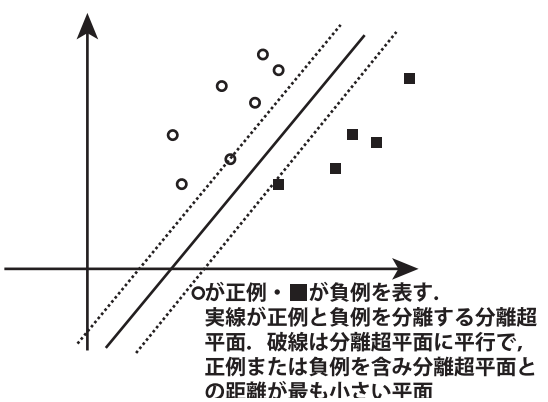

図 2　サポートベクターマシン（SVM）の概要

れぞれが，正例・負例いずれかのクラスに属しているとする．概要図を**図2**に示す．

図2において実線で示される，正例と負例を分離する超平面（分離超平面）を考える．一般にこのような分離超平面は一意には定まらないが，**図2**において破線で示される2つの超平面の距離が最も大きくなり，分離超平面が2つの超平面から等距離にあるものを最適解とする．

この時，n次元ベクトルWと定数bを適切に選ぶことによって，分離超平面は

$$(\mathrm{w}\cdot x)+b=0$$

という数式 を満たすn次元ベクトルX全体の集合として表すことができる．

また，破線で示される2つの超平面は，同じW，bを用いて，

$$(\mathrm{w}\cdot x)+b\pm 1$$

という数式 を満たすn次元ベクトルX全体の集合として表すことができる．

このとき，2つの超平面間の距離は

$$\frac{2}{\|\mathrm{w}\|}$$

で表されるので，これを最大にする，すなわち‖W‖を最小にするWおよびbを以下の制約条件のもとで求めればよい．

$$\begin{cases}(\mathrm{w}\cdot x_i)+b\geq 1 & (x_i\text{が正例})\\ (\mathrm{w}\cdot x_i)+b\leq -1 & (x_i\text{が負例})\end{cases}$$

このWおよびbを求める処理が「事例による学習」である．

実際には，超平面でこの2つを完全に分離することが不可能なように正例・負例が分布している場合が大半である．この場合は「ソフトマージン」という考え方を用いる．これは，**図3**に示すように本来の領域とは異なる位置にある事例ベクトルについて，破線で表される超平面との距離の2乗

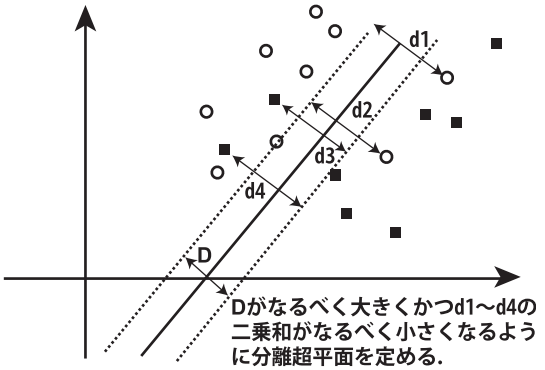

図3　ソフトマージンの概要

の総和が小さく，かつ‖W‖が小さくなるようにWを定める，というものである．

また，そのままでは超平面によって分離することが困難な場合，事例ベクトルを別のベクトル空間に適切に写像することによってうまく分離させることが可能になる場合も多い．

3つ以上のクラスに分離する場合も本質的には同様の処理である．この場合は，クラスごとに「そのクラスに属すか属さないか」で分離する．結果として「複数のクラスに属す」あるいは「どのクラスにも属さない」という場合もありうるが，その場合は分離超平面との距離などを用いて属するクラスを1つに決定する．

2）医学的知識の抽出への応用

SVMを構造化処理に用いる具体的な方法は，単語の情報（見出し語・品詞の種類・語の長さなど）をベクトルに変換し，このベクトルが特定の属性を持つかどうか（「部位」「所見」などに相当するかどうか）を求める際にSVMを用いる，というものである．

SVMを用いて特定の属性を持つ単語・フレーズを抽出するための言語処理ツールとして，奈良先端科学技術大学院大学松本研究室で開発された「YamCha」（http://chasen.org/~taku/software/yamcha/）がある．これは，形態素に分かれた文に含まれる各単語について，その単語および前後いくつかの単語の情報（日本語の場合は，単語そのもの・品詞・文字種 [漢字・カタカナ・ひらがな・英字・数字など]・単語の長さ・前後の単語の属性など）をベクトルに変換し，このベクトルに対

して SVM を用いることによって特定の属性を持つ単語・フレーズを抽出するというものである．

YamCha の使用例として，新聞記事を用いて学習を行い，これをもとに文章中から「地名」「人名」「組織名」「時間」「日付」などを抽出するという例がある[5]．

この例にならい，既存の自然文読影レポートを学習用データとして，処理の対象となる新しい自然文読影レポートから「部位」「基本所見」などの医学的知識を抽出する実験を行った．ここではまず YamCha の使い方について簡潔に述べる．

ステップ１：学習用データにタグ付け（属性の付与）を行う（基本的に手作業）．

一例として「前頭葉白質に高信号領域を認める」という文にタグ付けを行うと**表 1** のようになる．このようなルールで学習用データ内のすべての文にタグ付けを行う．

ステップ２：モデルを生成する（YamCha が自動で行う）．

タグ付けされた学習データを所定のルールでベクトルに変換し，これを用いて 2-1）で述べた分離超平面を計算する．タグの種類が多くなればもちろん分離超平面の数も増える．こうして得られた分離超平面の集合を「モデル」と呼ぶ．

ステップ３：モデルを用いて新しいデータにタグを与える（YamCha が自動で行う）．

処理の対象となる形態素解析済み文の各単語の情報をベクトルに変換し，これにステップ２で得られたモデルをあてはめて各単語に与えるタグを決定する．

表 1　タグの例

前頭葉	B-PART
白質	I-PART
に	0
高信号	B-CONDITION
領域	I-CONDITION
を	0
認める	B-CONCLUSION

PART…部位
CONDITION…基本所見
CONCLUSION…結語
0 …何もなし

特定の属性を持つフレーズの最初の語にはBを，
それ以外にはIを頭につける．

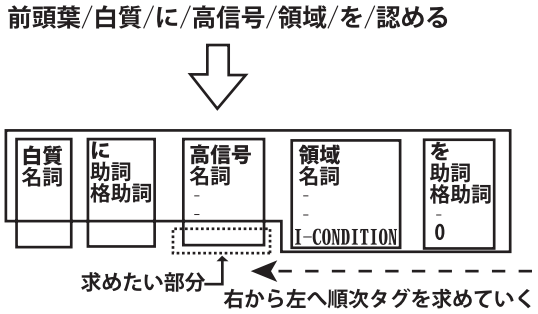

図 4　タグ付け処理のイメージ

ここで，アルゴリズムとしては「文の前から後ろに向かって処理を行う（タグを与える際に前の単語のタグ情報を利用する）」「文の後ろから前に向かって処理を行う（タグを与える際に後ろの単語のタグ情報を利用する）」の２種類があるが，日本語においてはその構造上，文の後ろから前に向かって処理を行う方が適切な結果が得られるので[5]，我々も文の後ろから前に向かって処理を行うことにする．

この処理のイメージを**図 4** に示す．

3．実　験

1）自然文読影レポートの文章構造と抽出する医学的知識

自然文読影レポートの文章構造および抽出すべき医学的知識は，そのレポートがどの部位について書かれたものであるか，あるいはモダリティが何であるかによって異なることが考えられるが，この実験では頭部 MR に関する読影レポートを対象とした．

まず，兵庫医科大学病院に提供していただいた頭部 MR に関するレポートについて文章構造を分析した．このレポートは計 270 件，１つのレポートが複数の文からなることもあるので文の数では 408 文である．その結果，文は大きく分けて「特徴記述文」「診断記述文」の２種類に分けられることがわかった．１つのレポート内に「特徴記述文」と「診断記述文」が別の文として含まれる場

表 2 「特徴記述文」「診断記述文」の例

特徴記述文
FLAIR像にて基底核に高信号域を認める． 基底核，深部白質に散在性の異常信号域が認められる． 左側脳室にHIAを認めます． 脳室に変形が見られます．
診断記述文
急性期のラクナ梗塞を疑う． 古い梗塞と考える． 炎症を第一に考えます． 小脳にクモ膜嚢胞が疑われる．
特徴記述文と診断記述文が同時に含まれる文
T2WIで右視床に小高信号域を認め，多発小梗塞と考えます． 大脳半球の深部白質に斑状の高信号域を認め，梗塞を疑います．

合が多いが，1 つの文に「特徴記述文」と「診断記述文」が共に含まれる場合もある．「特徴記述文」「診断記述文」の例を**表 2** に示す．

さらに，これらのレポートをさらに細かく分析し，「特徴記述文」「診断記述文」の文章モデルを求め，抽出すべき医学的知識は何であるかを調べた[2]．その結果を**表 3** に示す．

表 3 に示した，「撮影条件」「部位」「特徴」「基本所見」「結語（特徴記述文）」「診断」「結語（診断記述文）」を，今回の実験で抽出する医学的知識とした．

2）実験方法

実験の手順を以下に示す．

ステップ 1：学習用データおよび処理対象データを形態素解析する．

形態素解析には，奈良先端科学技術大学院大学松本研究室で開発された形態素解析器「茶筌（ChaSen）」（http://chasen.naist.jp/）を用いた．ただし，読影レポートには医学の専門用語が多く，「茶筌（ChaSen）」に付属の形態素解析用辞書だけでは正しく処理できないので，京都大学医学部附属病院の竹村匡正助手から提供いただいた用語集，医学中央雑誌刊行会発行の「医学用語シソーラス」，「ライフサイエンス辞書プロジェクト」作成の「ライフサイエンス辞書」（http://lsd.pharm.kyoto-u.ac.jp/）および財団法人医療情報システム開発センター作成の「標準病名マスター」（http://www.medis.or.jp/4_hyojyun/download/）に含まれる用語を辞書に追加した．

表 3 「特徴記述文」「診断記述文」から抽出すべき知識

特徴記述文				
文章モデル…「撮影条件」にて「部位」に「特徴」の「基本所見」を「結語」				
撮影条件	部位	特徴	基本所見	結語 (特徴記述文)
T2WI FLAIR像 MRA	基底核 脳室 大脳半球の 深部白質	散在性の 斑状の 5㎜程度の	高信号域 HIA 変形	認める 見られる
診断記述文				
文章モデル…「診断」を「結語」				
診断			結語(診断記述文)	
ラクナ梗塞 多発小梗塞 クモ膜嚢胞			疑う 考える	

学習用データには，前節でも用いた兵庫医科大学病院のレポートを，処理対象データには大阪大学附属病院の頭部 MR に関するレポートを用いた．この大阪大学附属病院のレポートは 74 件，文の数では 330 文であるが，このうち 68 文は「再検査お願いします」「MRA を order してください」などの「特徴記述文」「診断記述文」どちらの文章モデルにもあてはまらない定型的表現なので処理対象から除外したため，実質的な文の数は 262 文である．学習用データと処理対象データを別々の病院のレポートにして実験したのは，病院ごとのレポートの書き方の違いが「医学的知識」の抽出処理に及ぼす影響を加味するためである．

ステップ 2：2-2）で述べた方法で学習用データからモデルを作成し，これを用いて処理対象データにタグを付与する（「医学的知識」の抽出）．

ステップ 3：ステップ 2 で付与されたタグ（つまり，抽出した「医学的知識」）が正しいかどうかを検証する．

単語単位での正解，および文単位で見て正しく抽出できているものがどの程度あるかを求める．なお，文単位については，

・「抽出可能」（文全体にわたって「医学的知識」が正しく抽出されている）．

・「主旨は抽出可能」（一部，「医学的知識」が誤って抽出されている箇所があるが，言語処理の観点では主旨は正しく抽出されている）．

・「抽出不可」（それ以外）の 3 種類に分ける．「主旨は抽出可能」の例を**表 4** に示す．

なお，奈良先端科学技術大学院大学松本研究

表 4　「主旨は抽出可能」の例

「主旨は抽出可能」の例	そう判断した理由
「脳室周囲に強いT2領域を認める」という文において「強いT2領域」を「特徴+基本所見」として抽出するのが正解だが,「強い」が抽出できなかった場合.	「基本所見」として「T2領域」が抽出できていて,「T2領域」の意味は「T2強調像による高信号域」であることから「強い」が抜けていたとしても「主旨は抽出できている」と判断した.
「側脳室～第3脳室に拡大が見られる」という文において,「側脳室～第3脳室」を「部位」として抽出するのが正解だが,「側脳室」「第3脳室」を別々に抽出した場合.	部位として抽出すべき語である「側脳室」「第3脳室」を抽出できているので「主旨は抽出可能」と判断した.

室によって公開されているツールとして，「茶筌（ChaSen）」「YamCha」を含み，自然文から一気に形態素解析・タグ付け・文節区切り・係り受け解析までを行うことができる「南瓜（CaboCha）」（http://chasen.org/~taku/software/cabocha/）というツールがある．本実験ではこの「南瓜（CaboCha）」を用いてステップ 1 ・ステップ 2 の処理を一気に行った．

3）結　果

前節の実験の結果を**表 5** に示す．文単位で「抽出可能」「主旨は抽出可能」であったのは 262 文中 111 文（正解率 42.4％）であり，あまり高いとは言えない．

この理由は，前節でも少し述べた通り，病院ごとのレポートの書き方の違いによる影響と考えられる．具体例を挙げると，兵庫医科大学病院の文章は多くが「MRI にて脳に高信号を認める」のように「A にて B に C を認める」（A は「撮影条件」，B は「部位」，C は「基本所見」）という形になっている．したがって，兵庫医科大学病院の文章を学習データとして SVM を適用すると，この「A にて B に C を認める」という形の文章からは「医学的知識」を正しく抽出できるが，少し文章の表現が異なると，「医学的知識」を正しく抽出できないことが多くなる．

表 5　実験結果

文レベルで見た結果		単語レベルで見た結果	
処理の対象となる文の個数	262	処理の対象となる文に含まれる単語の総数	3,720
「抽出可能」であった文の個数	72（27.5%）	単語としての「正解」の個数	2,511（67.5%）
「主旨は抽出可能」であった文の個数	39（14.9%）	単語としての「不正解」の個数	1,209（32.5%）
「抽出不可」であった文の個数	151（57.6%）		

大阪大学附属病院の文章は「A にて B に C」「A：B に C を認める」「B に A で C を認める」などの様々な表現が使われているため，そのままでは「医学的知識」を正しく抽出できない例が多数出現したと考えられる．

そこで，SVM を適用する前に，処理対象データに使われている表現が学習データ中にも出現するように言い換え処理を行っておくと，「医学的知識」を抽出する精度の向上が期待できる．

4）改善案 1 ・機械的な言い換えを生成して学習用データに追加

前節で述べた言い換え処理の具体的な方法として考えられるのが，学習データの各文章に対し，意味は同じで助詞や結語表現が異なる文章を多数作成し，学習データ中に現れる表現の種類を増やす方法である．

そこで，学習用データとして用いている兵庫医科大学病院の各レポート文に対し，**表 6** に示すように，助詞・語順・結語部の表現などを一部変えた文を何種類か機械的に作成し，これらをすべて加えたものを学習用データとして，3-2）と同様の手順で大阪大学附属病院のレポートの処理を行った．

この結果を**表 7** に，このモデルの改良によって正しくタグ付け可能になった例を**表 8** に示す．文単位で「抽出可能」「主旨は抽出可能」であったのは 262 文中 136 文（正解率 51.9％）で，兵庫医科大学病院のレポートを単純にそのまま学習用データとして用いた場合に比べて精度向上が見られる．

5）改善案 2・処理対象となるレポートの言い換え検討

前節の処理によって精度は向上したが，それでも正しく処理できない例はまだ少ないとは言えない．

そこで，3-3）で述べた言い換え処理を前節よりもさらに拡張して考え，処理対象データの各文章を，意味はほぼ同じで表現が学習データとして

表 6　機械的に作成した文の例

元の文	MRIにて脳に高信号域を認め,脳梗塞を疑う.
作成した文	MRIにて脳に高信号域を認め,脳梗塞を疑う. MRIにて脳に高信号域を認める.脳梗塞を疑う. MRIにて脳に高信号域があり,脳梗塞を疑う. MRIにて脳に高信号域あり.脳梗塞を疑う. MRIにて脳に高信号域が認められ.脳梗塞を疑う. MRIにて脳に高信号域が見られ.脳梗塞を疑う. MRIにて脳に高信号域を指摘でき.脳梗塞を疑う. MRIにて脳に高信号域.脳梗塞を疑う. MRIにて脳に高信号域を認め,脳梗塞を疑う. 脳にMRIで高信号域を認め,脳梗塞を疑う. MRIにて脳に高信号域を認める.脳梗塞と思われる. MRIにて脳に高信号域を認め,脳梗塞と考える. MRIにて脳に脳梗塞を疑う高信号域を認める. MRIにて脳に,脳梗塞と考える高信号域を認める.

使われている文章に近くなるように言い換えるという方法を検討した.

具体的には，大阪大学附属病院のレポートのうち正しく処理ができなかったものについて，意味が大きく変わらない範囲で表現が学習データとして使われている文章に近くなるように言い換えを行い，言い換えた文に対して 3-4）で用いた学習データと同じものを学習データとして 3-2）と同様の手順で処理を行った.

この結果を**表 9** に，言い換えによって改善された例を**表 10** に示す.

表 7 の結果から，文単位で「抽出可能」「主旨は抽出可能」であった文は 262 文中 136 文であり，**表 9** の結果から，言い換えによって文単位で「抽出可能」「主旨は抽出可能」となった文は残りの 126 文中 95 文である．したがって，合わせると 262 文中 231 文（88.2%）が少なくとも適切に言い換えを行えば「抽出可能」「主旨は抽出可能」ということになる.

よって，多くの文は適切に言い換えを行えば正しく処理できると言える．逆に言えば，こうした言い換えを用いても正しく処理できない文は，読影レポート文として書き手の意図を正しく伝えるのにふさわしくない表現が含まれていると考えられる.

したがって，適切な言い換えに規則性を見つけ出すことができれば，機械的な対処で高い精度を確保することが可能となる．読影レポートを記述する際にいくつかのルールを定めることによって使われる表現の種類を絞り込めば，適切な言い換え規則を見つけることができると考えられるので，どういったルールを定めるのがよいかを今後の検討課題とする.

表 7　学習用データ追加後の実験結果

文のレベルで見た結果		単語レベルで見た結果	
処理の対象となる文の個数	262	処理の対象となる文に含まれる単語の総数	3,720
「抽出可能」であった文の個数	97 (37.0%)	単語としての「正解」の個数	2,724 (73.2%)
「主旨は抽出可能」であった文の個数	39 (14.9%)	単語としての「不正解」の個数	996 (26.8%)
「抽出不可」であった文の個数	126 (48.1%)		

表 8　正しくタグ付け可能になった例

正しくタグ付け可能になった例	正しくタグ付け可能になった理由(推定)
両側hypothalamusに浮腫によるT2延長領域.	「MRIにて脳に高信号域.」のように「結語」を省略した文を学習データに追加したため
両側前頭葉白質に数箇所梗塞と思われる5㎜程度の病変を認める.	「MRIにて脳に、脳梗塞と考える高信号域を認める.」のように,診断記述文が特徴記述文の中に含まれている文を学習データに追加したため
その周囲にはT2WIで淡い高信号域を伴っている.	「脳にMRIで高信号域を認め,脳梗塞を疑う.」のように,撮影条件が文の途中に入る文を学習データに追加したため

表 9　処理対象レポートの言い換えによる結果

3〜4の段階で「抽出不可」であった文の個数	126
言い換えによって「抽出可能」になった個数	71 (56.4%)
言い換えによって「主旨は抽出可能」になった個数	24 (19.0%)
言い換えによっても「抽出可能」「主旨は抽出可能」に至らなかった	31 (24.6%)

表 10　言い換えによって改善された例

元の文(「抽出不可」)	言い換えた文(「抽出可能」)
動脈瘤は明らかではない.	明らかな動脈瘤はない.
rtICAに頸部でstenosisがあるかもしれません.	rtICAに頸部でstenosisがありそう.
両側の内耳道を含め,SOLは指摘できません.	両側の内耳道を含めて,SOLは指摘できません.
軽度のischemic changeあります.	軽度のischemic changeあります.
明らかな狭窄や動脈瘤は指摘できません.	明らかな狭窄は指摘できません.動脈瘤も指摘できません.
これらの周囲にT2延長領域を認めます.	周囲にT2延長領域を認めます.

4. まとめと課題

本論文では，SVM を用いて読影レポート文か

ら「医学的知識」を抽出するための手法を示し，兵庫医科大学病院のレポート文を学習用データとして作成したモデルを用いて大阪大学附属病院のレポート文から「医学的知識」を抽出する実験を行い，文レベルで見ると 42.4％の正解率であることを確認した．さらに，SVM による「医学的知識」の抽出精度を高めるための工夫として，兵庫医科大学病院のレポート文に結語部分の何種類かの機械的な言い換え処理を行ったものを学習用データとして追加することにより，文レベルで 51.9％の正解率と，ある程度の精度向上が見られることを確認した．そして，「医学的知識」の抽出精度を高めるための別の工夫として，意味を大きく変えない範囲で表現を兵庫医科大学病院のレポート文に近づける言い換えを行うことにより，それまでの処理で正しく「医学的知識」を抽出できなかった大阪大学附属病院のレポート文についても，その多くは正しく「医学的知識」を抽出できることがわかった．

今後の課題は，上述した「意味を大きく変えない言い換え」を機械的な処理で行うことができる文の条件がどのようなものであるかを推定し，その条件をもとに読影レポート文を記述するルールを定めること，およびそのルールに従って記述された読影レポート文に対して「医学的知識」を抽出する精度がどの程度向上しているかを確認することである．

また，各レポート文（学習用データ・処理対象データともに）をあらかじめ「ブースティングアルゴリズム」[6] を用いるなどして「特徴記述文」「診断記述文」「特徴･診断の両方が記述された文」に分類しておき,「特徴記述文」「診断記述文」「特徴・診断の両方が記述された文」で別々に 3-2）のような実験を行えばさらなる精度向上の可能性がある．

謝　辞

本研究を実施するにあたって奈良先端科学技術大学院大学情報科学研究科の松本裕治教授から有益なアドバイスを多数いただきました．ここに感謝の意を表します．

また，レポートを提供していただいた兵庫医科大学病院中央放射線部の中尾宣夫教授，および大阪大学医学部附属病院放射線部の中村仁信教授に感謝の意を表します．

また，日本医療情報学会の以下の先生方に多大なるご協力をいただきました．ここに感謝の意を表します．

兵庫県立大学大学院 応用情報科学研究科　稲田 紘教授

京都大学医学部附属病院 医療情報部　黒田知宏講師

京都大学医学部附属病院 医療情報部　竹村匡正助手

大阪大学歯学部 口腔総合診療部　玉川裕夫助教授

関西医科大学 病院医療情報部　仲野俊成講師

大阪市立大学医学部 医療情報部　朴　勤植助教授

兵庫医科大学 医療情報部　平松治彦講師

大阪大学医学部附属病院 医療情報部　松村泰志助教授

兵庫医科大学 医療情報部　宮本正喜教授

（五十音順）

参 考 文 献

1）笹井浩介．利用者の意図が理解できるデータベース検索システムの開発．月刊ファームステージ 2004；**4**：33-40

2）川上洋一，安永　晋，笹井浩介，他．症例データベースから抽出した医学的知識のレポーティングシステムへの応用．医療情報学 2005；**25**：962-965

3）安永　晋，川上洋一，笹井浩介．文章構造化技術の医用データベースへの応用．コニカミノルタテクノロジーレポート vol. 2　2005：113-118

4）Vladimir N Vapnik. Statistical Learning Theory. John Wiley & Sons, 1998

5）山田寛康，工藤　拓，松本裕治．Support Vector Machine を用いた日本語固有表現抽出．情報処理学会論文誌 vol.43　2002 ; **1** : 43-53

6）工藤　拓，松本裕治．半構造化テキストの分類のためのブースティングアルゴリズム．情報処理学会論文誌 vol.45 2004 ; **9** : 2146-2156