医学用語辞書で学習した分類器による放射線読影レポート用語の分類

田中昌昭，†竹内孔一
川崎医療福祉大学，†岡山大学大学院
mtanaka@mw.kawasaki-m.ac.jp, †koichi@cl.cs.okayama-u.ac.jp

1. はじめに

放射線読影レポートの所見部分に記載された文章から病態所見，形態的・画像的特徴，臓器・部位等を抽出して，後利用のできる形に構造化するには用語の抽出と分類，つまり意味クラスの付与が必要となる．たとえば「胆嚢頚部や総胆管内には欠損像が認められ，結石と思われます」という所見部分から「胆嚢頚部，総胆管内」，「欠損像」，「結石」といった用語を抽出し，それぞれ部位，画像的特徴，病態所見といった意味クラスを付与しなければならない．

固有表現抽出(NER: Named entity extraction)として知られるこのような処理は自然言語処理分野ではよく研究されており，様々な手法が提案されている[7]．しかし，そのほとんどが正解付きコーパスを利用して機械学習を行い，それを Gold standard に用いて評価を行うといった手法をとっている．たとえば，バイオインフォマティクスの領域では OHSUMED, GENIA といった文献共有による大量タグ付コーパスがコミュニティの努力によって維持管理されており，これらを利用して異なる手法の客観的な比較を行うことが可能になる[11]．

ところが放射線読影レポートなどの医療文書は個人情報であるため，新聞記事や論文アブストラクトなどと違って正解付きコーパスが公開されることはなく，今後も利用できる状況になるとは限らない．また，用語辞書もあるにはあるが，流通しているのはいずれも汎用的な医学用語辞書ばかりで，放射線分野に特化した，しかもレポートに用いられる用語を集めたような辞書は筆者の知る限りない．このような状況で，正解付きコーパスを人手で構築するには多くの労力を要する．そこで，本研究では放射線読影レポートの所見部分の構文情報から用語候補の抽出を行い，既存の医学用語辞書を訓練コーパスに用いて学習した分類器によって用語の分類を試みた．

2. 対象データと方法

医学用語の多くは複合語であるという特徴を利用して，用語の構成要素を手掛かりに意味クラスを判定する分類器を構築した．

医学用語辞書[1]に収載されている約 18 万語の日本語見出し語を形態素解析器によって構成要素に分割した．個々の用語には解剖や疾患など 19 種類の分類種別が割り当てられている．この分類種別を用語に付与すべき意味クラスとし，用語の構成要素を素性として用語の分類種別を判定する分類器を構築した．

次に，放射線読影レポートから検査種別が MR のレポートを無作為に 1,000 件抽出して構文解析を行い，得られた 25,943 個の文節のうち，用言以外の 19,162 個の文節から助詞を削除した語幹部分（異なり総数 5,360）を分類対象用語として抽出した．

3. 分類器

分類器として辞書ベース(DB: Dictionary based)，naïve Bayes(NB)，最大エントロピー法(ME: Maximum Entropy Method)，サポートベクタマシン(SVM: Support vector machine)，条件付確率場(CRF: Conditional random fields)の 5 つの機械学習手法を用い，これらの分類器を複数足し合わせた投票(Vote)によって判定を行った．

3-1. 辞書ベース (DB)

用語の構成要素を医学用語辞書と照合して分類を行う．これを本研究では辞書ベースの分類と呼ぶ．具体的には，分類対象の用語に完全あるいは部分一致する医学用語を抽出し，その種別によって用語の種別を判定する．たとえば，分類対象の用語が「右心房静脈洞」の場合，用語を構成する「右」，「心房」，「静脈洞」の 3 つの部分文字列のうち，「心房」と「静脈洞」は解剖用語であるが，「右」は解剖用語ではない．この場合，解剖用語への帰属率を 2/3 と定義する．このように，閾値を導入し，ある種別への帰属率が指定した閾値以上になる場合に限ってその種別を分類結果として出力し，そうではない場合は分類不能とした．なお，本研究では閾値を 0.5 とした．

3-2. Naive Bayes (NB)

用語の構成要素を素性として利用して分類を行う．構成要素として，医学用語を形態素解析器（MeCab[2]を使用）にかけて得られた形態素（表層）を用いる．ただし，形態素解析器の辞書はデフォルトのものを使い，ユーザ辞書は用いない．今，用語 w を形態素解析して得られた形態素を $m_1,\cdots,m_n$ とすると，与えられた用語 w の種別 c は

$$c = \arg\max_{c'} P(c' \mid w)$$

$$= \arg\max_{c'} P(c') \prod_{k=1}^{n} P(m_k \mid c')$$

によって求める．種別の事前確率 $P(c)$ や条件付き尤度 $P(m \mid c)$ などのモデルのパラメタは学習データより求めた．

3-3. 最大エントロピー法 (ME)

最大エントロピー法は，与えられた制約のもとで，エントロピーを最大にするような条件付き確率分布を求める手法である．naive Bayes と異なり，独立性の仮定を必要としないので，素性を自由に選べるという利点があるが，本研究では，素性として用語 w の形態素 $m_1,\cdots,m_n$ のみを用いた．用いた素性関数は以下のようなものである．

$$f_k(m_1^n,c)=\begin{cases}1 & m_1^n \in L(c)\\ 0 & \text{otherwise}\end{cases}$$

ここで，$L(c)$ は訓練コーパス中の種別 c に属する医学用語の構成要素を集めたものである．なお，最大エントロピー法による分類ソフトウェアとして OpenNLP MaxEnt[3]を用いた．その際，iteration 回数としてデフォルト値の 100 を用いた．

3-4. サポートベクタマシン (SVM)

SVM は，次元拡張された素性空間内に学習データから最大マージンが得られるような超平面を求め，それによって未知データを 2 値分類する機械学習法である．

本実験では，素性空間として用語 w の形態素 $m_1,\cdots,m_n$ の出現頻度を要素とするベクトル空間を用いて用語の分類を行った．SVM は 2 値分類器であるため，そのままでは医学用語を 19 種類の種別に分類することはできない．そこで，SVM を多値分類問題に適用できるように拡張された SVMmulticlass[4]を用いた．なお，分類実験では線形カーネルを用い，C (regularization parameter) の値は予備実験によって求めた最適値 10,000,000 を使った．

3-5. 条件付確率場 (CRF)

CRF は最大エントロピー原理に基づいた識別モデルで，最大エントロピー法の出力が構造を持たない（即ちスカラー量である）のに対して，CRF は構造を持った出力を得ることができる．そのため，CRF は入力単語系列から品詞系列を出力する形態素解析や，入力単語系列から固有表現を抽出するタスクなど，観測系列のタギングに用いられる．

本実験では入力系列として医学用語を形態素解析して得られた形態素（表層）を与え，出力系列として用語の種別（の BIO タグ）を求めるタスクとして CRF を用いた．これは本来の CRF の使用法とは異なるが，CRF は本質的に最大エントロピー法と同じ手法なので，用語の分類のように出力が構造を持たないタスクに対してどの程度の分類性能を示すかを調べてみるために本実験を行った．なお，CRF には CRF++[5]を用い，素性テンプレートには Window サイズ±2 の表層およびそれらを合成した Unigram 素性と Bigram 素性を用いた．

4. 評価方法

最初に，医学用語辞書を用いて訓練した 5 つの分類器の性能評価を 10-folds cross validation で行った．つまり，医学用語辞書を 10 分割し，分割されたデータセットのうち 9 つを使って訓練した分類器を用いて残りの 1 つのデータセットの検証を行った．これをすべてのデータセットに対して 10 通りの分類実験を行い，その結果を総合して分類性能評価指標を計算した．分類性能評価指標として再現率，適合率，F 値を計算して各種手法の比較を行った．

次に，放射線読影レポートから抽出した用語を 3 種類の辞書（医学用語辞書[1]，医学用語シソーラス，最新解剖学用語集[6]）と完全文字列マッチングを行って，辞書のカバレッジを調べた．これはベースラインとして評価の際に用いた．

最後に，放射線読影レポートから抽出した 5,360 語の候補用語から分類器の訓練に用いた医学用語辞書に収載されている 820 語を除いた 4,540 語からランダムに 1,000 語を抽出して人手による分類種別（意味クラス）の付与を行い，これを Gold standard として各分類器の再現率・適合率・F 値を計算して性能評価を行った．

5. 結果

5-1. 医学用語辞書で訓練した分類器の性能

医学用語辞書を用いて学習した各分類器の性能を 10-folds cross validation によって評価した結果を表 1 に示す．表には全体の正解率（分類器の正解数を用語総数で割ったもの）に加えて解剖，疾患の 2 種類の分類種別について計算した再現率(Recall)と適合率(Precision)を示している．

全体の正解率では，CRF が 74.5%と最高の性能を示した．これに対して辞書ベースの正解率は僅かに 24.4%で，5 つの分類器の中では最低であった．種別ごとに分類性能を見た場合，解剖用語についてはやはり CRF が最高 (F=0.829)で，その後 SVM, ME, NB と続いたが，辞書ベースでも F=0.623 と比較的高い性能を示した．一方，疾患用語では，CRF, SVM, ME, NB の F 値が 0.8 を超えていたのに対して辞書ベースは 0.265 と他の手法に比べて極めて低かった．

表 1. 10-folds cross validation による性能評価

手法		DB	NB	ME	SVM	CRF
解剖	Recall	60.0%	78.6%	83.2%	85.0%	83.8%
	Precision	64.7%	79.2%	79.2%	78.8%	82.1%
	F	0.623	0.789	0.812	0.818	0.829
疾患	Recall	15.9%	81.9%	84.7%	86.3%	86.2%
	Precision	80.1%	79.1%	84.3%	86.3%	86.8%
	F	0.265	0.805	0.845	0.863	0.865
全体の正解率		24.4%	65.2%	68.4%	70.2%	74.5%

5-2. 辞書との完全文字列マッチによる評価（辞書のカバレッジ）

３種類の辞書（医学用語辞書，医学用語シソーラス，最新解剖学用語集）と完全文字列マッチングを行った結果を表 2 に示す．

表 2. 各種辞書によるカバレッジ

辞書	異なり語レベル	語レベル
医学用語辞書	820/5,360（15.3%）	4,996/19,162（26.1%）
医学用語シソーラス	349/5,360（6.5%）	2,088/19,162（10.9%）
最新解剖学用語集	454/5,360（8.5%）	1,926/19,162（10.1%）

語レベルでは 10～26%，異なり語レベルでは 6～15%という結果であった．

5-3. レポートから抽出した用語の分類

医学用語辞書を使って訓練した5つの分類器を使って放射線読影レポートから抽出した用語を分類した結果を表 3 に示す．表中の Vote は 5 つの手法の多数決によって種別を判定した結果である．

表 3. レポートから抽出した用語の分類結果

手法	解剖用語			疾患用語		
	再現率	適合率	F	再現率	適合率	F
DB	64.3%	78.7%	0.708	11.5%	70.0%	0.198
NB	71.9%	83.4%	0.772	65.9%	59.4%	0.625
ME	84.8%	83.9%	0.844	75.8%	65.1%	0.701
SVM	89.0%	78.9%	0.836	79.1%	60.8%	0.687
CRF	72.5%	83.8%	0.777	73.6%	47.3%	0.576
Vote	87.9%	83.0%	0.854	78.6%	66.8%	0.722

解剖用語については，再現率は SVM による分類が 89.0%で最も高く，適合率は ME が 83.9%で最高であった．F 値で総合性能を見た場合は「多数決」が 0.854 で最も高性能であった．解剖用語の分類性能で興味深いのは辞書ベースでも F 値で 0.708 もの分類性能を示している点である．

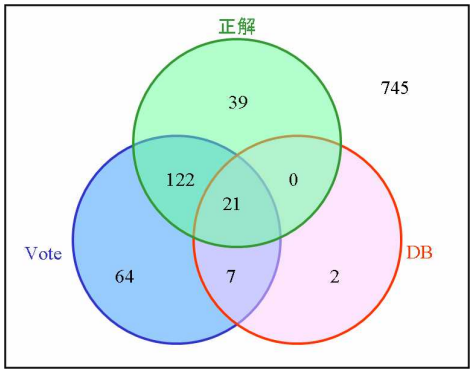

図 1. 辞書ベース(DB)と多数決(Vote)の比較（疾患）

一方，疾患用語については，再現率は SVM が 79.1%で最も高く，適合率は辞書ベースが 70.0%で最も高かった．F 値で総合性能を見た場合は「多数決」が 0.722 で最も高性能であった．特筆すべきは，疾患用語の分類では，辞書ベースの適合率が他の手法よりも高い値を示しているという結果である．しかし実は，これは見せかけだけの話で，図 1 に示すように辞書ベースは，True Positive，False Positive ともに件数が少なく（つまり，分類器の疾患用語検出能力が低く）取りこぼしが多いだけのことである．辞書ベース以外の手法はいずれも適合率よりも再現率が高い．つまり，疾患用語を検出する能力が高い．反面，それだけ False Positive も多くなる．

6. 関連研究

今井らは放射線読影レポートの文型分類を行い，（[部位] ニ | * | [所見] ヲ | [認める]）等の 22 種類の格フレームを構築し，325 文のテストセットから「部位・所見・主張句」を抽出して再現率＝45.6%，適合率＝97.6%を得たと報告している[8]．しかし，この手法では複文や長い文章はルール化が困難とも述べている．

Jin 等は CRF を用いて Medline のアブストラクトから悪性腫瘍の表現を抽出する実験を行い，再現率=81.8%，適合率=85.1%，F=0.834 を得たと報告している[9]．また，CRF の入力データに悪性腫瘍に固有な素性を加えたところ，再現率は 84.6%と向上したものの，適合率は 83.1%と下がり，F 値も 0.838 と殆ど変わらず，分野固有の素性は悪性腫瘍表現の抽出に寄与しなかったと述べている．彼らはまた悪性腫瘍用語リストとの完全文字列一致による悪性腫瘍表現の抽出も試みており，正しく抽出できたのは 42.1%程度だと報告している．英語と日本語の違いはあるが，本研究でも辞書との完全文字列一致で抽出できるのは 1～2 割程度であった．

Tsuruoka 等は，CRF など機械学習手法による固有表現抽出では単語は抽出できてもその ID を同定できないとして，辞書ベースのアプローチをベースとした手法を考案し，GENIA コーパスを対象として行ったタンパク質表現の抽出実験で再現率=67.2%，適合率=73.5%，F=0.702 を得たと報告している[10]．彼らが考案した手法の特徴は，辞書との完全文字列マッチを行うのではなく，edit distance と呼ばれる類似度を指標とした近似文字列検索によってタンパク質表現の候補を抽出し，それを naïve Bayes 分類器でフィルタにかけ適合率を向上させている点である．これに対して本研究のアプローチは構文情報から文節を抽出し，得られた用語候補を 5 つの機械学習手法で分類している点が異なっている．

7. 考察

表3に示す分類結果はレポートから抽出した用語が学習に用いた医学用語辞書にはなかったデータ，つまり，真に未知のデータに対する分析結果である．学習データに対する cross validation の結果（表 1）と比較すると，容易なタスク（解剖用語の分類）ではさほど分類性能に違いは見られなかったが，困難なタスク（疾患用語の分類）では 5 つの手法とも軒並み性能を下げており，しかも，性能にバラツキが見られた．手掛かりが少ないことに加え，各学習モデルが過学習に陥っている可能性があり，レポート中に出現する疾患用語の特徴を十分に捉え切れていないものと考えられる．

これに対して分類器を複数組み合わせて多数決による判定を行うと，解剖用語では F=0.854，疾患用語では F=0.722 と分類性能が安定し，各学習モデルのバラツキをうまく修正することができた．以下に具体例を示してこの理由を考察する．

表 4. 疾患用語の誤分類例

語幹部	DB	NB	ME	SVM	CRF	Vote
34×28mm大	分類不能	集団	症状	解剖	疾患	分類不能
23X37mm大です	分類不能	集団	症状	解剖	疾患	分類不能
24mm大	分類不能	症状	症状	解剖	疾患	症状
28x18x16mm大	分類不能	集団	症状	解剖	疾患	分類不能
25x26x54mm大	分類不能	集団	症状	解剖	疾患	分類不能

表 4 は，疾患用語の分類において CRF が誤分類した事例の一部を示したものである．CRF は「数字, X, mm，大」などのトークンから構成されるサイズ表現を疾患用語に誤分類している．分類器の学習に用いた医学用語辞書の見出し語に含まれる「大」の出現割合は，表 4 の分類結果に現れる 4 つの種別の中では疾患用語が最も低かった．それにも拘わらず，CRF が疾患用語と判定した理由は「リンパ節腫大」のように「・・・大」で終わる疾患用語の出現割合が最も高かったからである．CRF は他の学習モデルと違ってトークンの出現順を学習している．一方，SVM はサイズ表現を解剖用語に分類している．「大」の出現割合に加えて「X」の出現割合が高い解剖用語の特徴を SVM は学習しているものと考えられる（解剖用語よりも「大」の出現割合が高い症状用語や集団用語には「X」というトークンは現れていなかった）．また，これらのサイズ表現を，naïve Bayes は集団用語に，ME は症状用語に分類している．「大」の出現割合が最も大きいのは症状用語であった．それにも拘らず naïve Bayes が集団用語に分類した原因はゼロ頻度問題を回避するために用いた尤度の推定式にあった．訓練コーパス中に出現しないトークン（あるいは低頻度トークン）に対しては，訓練コーパス中のサイズが小さいクラスに有利なバイアスがかかる．一方，ME は naïve Bayes と違って訓練データの素性の組合せを学習しており，「大」に関しては頻度の面でも素性の組合せの面でも ME に有利に働いていた．

このように各学習モデルにはそれぞれ素性の捉え方に特徴があり，その結果，性能にバラツキが生じたが，それらを組み合わせることによって個々の学習モデルが相互に補い合い，安定した性能が得られたものと考えられる．

本研究で提案する手法の問題点は，放射線読影レポートから抽出した用語の分類，つまり，意味クラスの特定はできても，その ID の同定まではできないという点である。これについては，分類済みの用語に対して分野オントロジーを適用するなど上位層での解決が必要となるが，今後の課題として残された．

今回はレポートから抽出した用語を入力データとして，医学用語辞書によって学習した分類器で用語を分類するという方法を提案したが，今後は，これを拡張することで用語抽出を行う方法を構築したい．たとえば，その応用として，提案する手法を放射線読影レポート等の自動アノテーションに利用することによりタグ付きコーパスの効率的な構築に利用できるものと考えている．また，素性に分野固有の知識を加える手法としての応用が期待される．

参考文献

[1] 医学用電子化 AI 辞書研究会, 25 万語医学用語大辞典, 日外アソシエーツ株式会社，1991.

[2] MeCab: Yet Another Part-of-Speech and Morphological Analyzer, http://mecab.sourceforge.net/

[3] OpenNLP MaxEnt: http://maxent.sourceforge.net/

[4] SVMmulticlass: Multi-Class Support Vector Machine. http://svmlight.joachims.org/svm_struct.html

[5] CRF++: Yet Another CRF toolkit, http://crfpp.sourceforge.net/

[6] 最新解剖学用語集．2007 年 8 月 1 日改訂版．http://web.sc.itc.keio.ac.jp/anatomy/TA/TA-contents.html

[7] Zweigenbaum P, Demner-Fushman D, Yu H, Cohen KB; Frontiers of biomedical text mining: current progress, *Brief Bioinform*, Vol.8, No.5, pp.358-75 (2007).

[8] 今井健，小野木雄三；格フレームを用いた放射線読影レポートの文型分類と所見抽出. *医療情報学 24 回連合大会論文集*, pp.800-801 (2004).

[9] Jin Y, McDonald RT, Lerman K, Mandel MA, Carroll S, Liberman MY, Pereira FC, Winters RS, White PS; Automated recognition of malignancy mentions in biomedical literature, *BMC Bioinformatics 2006 Nov 7*;7:492.

[10] Tsuruoka Y, Tsujii J; Boosting Precision and Recall of Dictionary-Based Protein Name Recognition, *Proceedings of the ACL-03 Workshop on Natural Language Processing in Biomedicine*, pp.41-48 (2003).

[11] Collier N, Nobata C, and Tsujii J; Extracting the Names of Genes and Gene Products with a Hidden Markov Model, *Proceedings of 18th COLING*, pp.201-207 (2000).