小規模な用語リストを利用した画像読影レポートからの用語抽出

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

電子カルテの普及に伴い、病院情報システムには退院時サマリや画像読影レポートなど自然言語で記述された医学文書が大量に蓄積されつつある。しかしながら、医学用語など専門用語を多用する医学文書の自然言語処理に関する研究はまだ緒についたばかりであり、電子的に蓄積された医学文書の利用は 1 次利用にとどまっているのが現状である。このような背景の下に我々は画像読影レポートから部位や症状などの専門用語を抽出する試みを行ってきた[1,2]。いわゆる固有表現抽出タスクと呼ばれるこうした要素技術は新聞記事を対象とした汎用的な領域や MEDLINE などの医学生物学領域の文献データベースを対象としたバイオロジ分野では盛んに研究され多くの成果を挙げているが、医学分野は扱う情報の秘匿性から大規模なコーパス開発による公開研究に馴染まず、研究者はゼロからコーパスを構築することを強いられるのが常である。なかでも多くの労力と専門的な知識を要するのが正解ラベルを付与するアノテーション作業である。そこで、この負担を少しでも軽減することを目的として、入手可能な小規模な用語リストを用いて人手に頼ることなくコーパスに正解ラベルを付与する手法を考案した。

先行研究として、Collins 等はたった 7 個の初期ルール（seed）に基づいて、スペル素性と文脈素性を交互に利用して決定リスト（Decision List）を作成し、それを用いて用語にラベル（Person, Organization and Location）を付与した[3]。また、Torii 等は GENIA コーパスから用語のタイプを推定する決定リストを自動獲得し、それを用いてタンパク質や DNA などバイオロジ領域の固有表現を分類した[4]。彼らもまた用語に含まれる単語やサフィックスなどの内部素性と共起単語やその出現位置などの外部素性（文脈素性）を利用している。本研究でもこれらの先行研究と同様に表層素性と文脈素性を用いて決定リストを作成し、それを用いて読影レポートに出現する用語の分類を行ったが、決定リストを作るための学習データの初期集合として入手可能な小規模な用語リストを用いた点と学習データ増強のために行うブートストラッピングの各ラウンドにおいて最適な素性セットを自動獲得した点が異なっている。

提案手法の有効性を確認するために教師付き学習手法である最大エントロピーモデルと比較したところ、最大エントロピーモデルでは $F = 0.668$ であったのに対して提案手法では $F = 0.856$ となり、本手法の有効性が確認された。

2. 材料と方法

2.1 用語候補の抽出

画像読影レポートから検査種別が MR のものをランダムに 500 件抽出し、得られた 2,041 のセンテンスを 2,854 の自立した節に分割した。次いで、節を構文解析（JUMAN+KNP）して 12,462 のチャンク（文節）を得、それらのうち用言ではない 9,371 文節を抽出し、その内容語を用語候補として意味属性を付与する対象とした。たとえば「L4 椎体は椎体の上半分が T1 強調像で低信号を呈しており、急性期の圧迫骨折が疑われます。」という文を「L4 椎体は椎体の上半分が T1 強調像で低信号を呈しており」と「急性期の圧迫骨折が疑われます」に分割し、構文解析によって「L4 椎体は」「椎体の」「T1 強調像で」などの文節を取り出し、助詞などの機能語を除去して「L4 椎体」「椎体」「T1 強調像」などの内容語を用語候補として抽出した。なお、日付表現記述（103 事例）とサイズ記述表現（113 事例）については正規表現を用いて抽出したので、これらを除く 9,155 事例（実際には空白文節が 1 つあったので、これを除く 9,154 事例）を意味属性付与の対象とした。

こうして得られた用語候補 9,155 事例に対して次節で述べる方法で意味属性の付与を行った。

2.2 用語の意味属性と使用する素性

抽出した用語候補に表 1 に示す意味属性を付与するために表 2 に挙げた 9 種類の素性を定義した。

表 1. 用語に付与する意味属性とその例

意味属性	LABEL	例
部位記述表現	BODYLOC	椎間腔、椎体終板、右付属器領域、左卵巣
問題記述表現	PROBLEM	狭小化、不整、炎症、圧迫骨折、悪性腫瘍
撮影記述表現	IMAGING	MRA、T1強調像、T2WI、脂肪抑制像
信号記述表現	SIGNAL	輝度変化、T1低信号、高信号、異常信号
造影記述表現	CONTRAST	SPIO造影、ガドリニウム造影、早期濃染
サイズ記述表現	SIZE	2cm大、6mm程度、径62×69mm
日付記述表現	DATE	04/10/1、1/28、2004.1.9、11月22日

表 2. 使用する 9 種類の素性とその例

#	Feature	Description	Example
f1	構文スケルトン1	用語が含まれる節の構文パタン	～ニ格～ガ格～
f2	構文スケルトン2	f1において用語が出現する位置を＊で示したパタン	～ニ格＊ガ格～
f3	格	用語に付随する格	ガ格
f4	共起単語	用語が含まれる節において用語と共起する他の用語	L1、認められます
f5	述語	用語に付随する格と用語が含まれる節の述語	ガ格@認める
f6	否定形	用語が含まれる節が否定形であるか否か	FALSE
f7	係り先	用語に付随する格と係り先	ガ格@認める
f8	接尾辞	用語を構成する要素（形態素）の最後に位置するもの	骨折
f9	形態素	用語を構成する要素（形態素）	圧迫、骨折

表 2 に示す素性のうち f1 から f7 までは用語の役割を示す格情報や共起する単語などの文脈素性で、f8 と f9 は用語を構成する要素（形態素）や接尾語などの表層素性である。Example 欄には、節「L1 に圧迫骨折が認められますが」に出現する用語「圧迫骨折」に対する素性値を示す。実際の用語分類ではこれらの素性の様々な組み合わせ（これを素性セットと呼ぶ）の

中から最適な素性セットを選択しながら処理を進めた。

2.3 決定リストを用いた用語の分類

まず、意味属性が C に属する N_C 個の学習データを $t_k(k=1,\cdots,N_C)$ とする。次に、用語 t の素性 f_i の値 v を求める関数 $v=g_i(t)$ を定義する。ここで、$i\in\{1,\cdots,K\}$ は素性の種類を表す添え字である。たとえば素性として用語の接尾語を考えた場合、用語 $t=$”脳梗塞”に対して $g_i(t)$ の値 v は”梗塞”となる。

次に、学習データ中の用語インスタンス集合から素性 f_i の確率分布を次式によって推定する。

$$\hat{P}_i(v\mid C)=\frac{Count(v,C)+\alpha}{\sum_v Count(v,C)+\alpha|v|} \quad \cdots (1)$$

ここで、$Count(v,C)$ は素性 f_i の値が v である用語インスタンスの出現頻度で、

$$Count(v,C)=\sum_{k=1}^{N_C}1[v=g_i(t_k)]$$

によって計算する。なお、$1[\cdots]$ はユニット関数で、引数内の条件が成立した場合に 1 を、そうでない場合は 0 をとる。また、α はゼロ頻度問題を解消するためのスムージングパラメタ、$|v|$ は v の異なり数である。こうして推定した確率分布は学習データを使って得られる決定リストと見ることができ[1]、これを用いて前節で求めた用語候補 t_j に対して、その用語の意味属性が C である確度を次式で計算することができる。

$$L_C(t_j)=\sum_{i=1}^{K}\log\hat{P}\left(g_i(t_j)\mid C\right) \quad \cdots (2)$$

レポートから抽出した全用語候補に対してこの確度を求め、その降順に整列すると、用語候補を意味属性が C である確度の高い順に並べたリストが得られる。このリストを用いれば、適当な閾値を設けることにより、意味属性が C である用語を抽出することが可能となる。こうして得られた用語を次のラウンドの学習データとしてこれまでの手順を繰り返せば、また新しいリストが得られる。つまり、最初に小規模の用語リスト（初期 seed）を用意して、この手順を繰り返すことにより（ブートストラッピング）、初期 seed に与えた用語と同じ意味属性を持つ用語を自動的に収集することが可能となる。以上述べた手順を図 1 に示す。

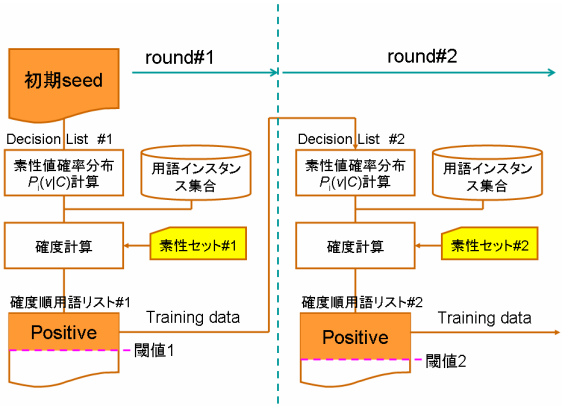

図 1. 決定リストを用いた用語分類手順

使用する素性 $\{f_1,\cdots,f_K\}$ は繰り返しの各ラウンドで異なる組み合わせを用いてもよい。また、確度順に整列した用語リストから次のラウンドに用いる学習データを抽出する際の閾値をどうするかが問題となる。各ラウンドで用いる最適な素性セットと閾値の決定方法については次節で説明する。

2.4 初期 seed と最適なパラメタの決定

図 1 に示す手順は初期学習データを必要とするが、人手による正解付与はコストが高くつくので避けたい。そこで、医療分野で開発されている各種の専門用語集を利用することにした。具体的には、意味属性が部位記述表現の初期 seed として、用語インスタンス集合のうち最新解剖学用語集に収載されている 439 事例（異なり語で 162 語）に加え、レポートに頻出する C3 や L4 などの頚椎や腰椎の略語を加えた 603 事例を用いた。また、意味属性が問題記述表現の初期 seed として、用語インスタンス集合のうち ICD10 対応標準病名集に収載されている 257 事例（異なり語で 89 語）を用いた。これ以外の意味属性については代表的な用語を手作業で拾い初期 seed とした。

次に、図 1 に示す手順の各ラウンドで用いる素性セットと閾値の決定方法について説明する。上述したように初期 seed は分類対象となる用語インスタンス集合の中に混在している。そこで、もし用語インスタンス集合中の正解データの中にこの初期 seed が均等に分布していると仮定した場合、初期 seed をマーカとして正解データの振る舞い（確度順用語リスト内での順位）を推定できるかもしれない。このアイデアに基づいて、本来ならば正解データによって算出すべき性能指標の代わりに初期 seed から算出した性能指標を判定基準に用いて、各ラウンドの最適なパラメタ（素性セットと閾値）を決定した。具体的には各ラウンドの確度順用語リストから、次式によって初期 seed の

[1] Collins[3]や Yarowsky[5]は(1)式ではなく $\hat{P}(C\mid v)$ を計算している。素性が互いに独立であるという仮定をおけば Bayes の法則を用いて(1)式から $\hat{P}(C\mid v)$ を計算することができるが、その場合事前確率 $\hat{P}(C)$ が必要となる。本研究では意味属性ごとに別々の決定リストを用いて binary 分類問題としているので $\hat{P}(C)$ が不要となり(1)式をそのまま用いた。

F 値を計算し、これを指標として最適なパラメタを選択した。

$$F_{IS} = \frac{2R_{IS}P_{IS}}{R_{IS} + P_{IS}}$$

$$R_{IS} = \frac{\text{Positiveデータ中の初期seedインスタンス数}}{\text{初期seedの総インスタンス数}}$$

$$P_{IS} = \frac{\text{Positiveデータ中の初期seedインスタンス数}}{\text{Positiveデータの総数}}$$

ここで Positive データとは図 1 に示す確度順用語リストにおいて、与えられた閾値以上の確度を持つ用語インスタンスのことである。以上をまとめると次のようなアルゴリズムになる。

- *imax* 個の素性セットを用意する。
- 初期 seed を作成して学習データとする。
- 用語インスタンス集合に含まれる初期 seed にマーカを設定する。
- Loop for $k = 1,...,kmax$
  - 学習データを使って素性値確率分布(Decision List)を計算する。
  - 用語インスタンス集合の個々のインスタンスに対して確度を計算する。
  - Loop for $i = 1,...,imax$
    - ✧ i 番目の素性セットを用いて確度順用語リストを作成する。
    - ✧ 確度順用語リストを使って初期 seed マーカによる最大 F_{IS} 値とそれを与える閾値（Positive データ数）を求める。
  - 試した素性セットのうち初期 seed マーカによる最大 F_{IS} 値を与える閾値がもっとも大きいものを最適な素性セットとして選択する（これを閾値最大基準と呼ぶ）。
  - 最適な素性セットを使って作成した確度順用語リストの Positive データ（確度が上記で求めた閾値を超える用語インスタンス）を次のラウンドの学習データとする。

3. 実験と結果

前章で述べた方法を用いて意味属性が部位記述表現の用語分類実験を行った。読影レポートから抽出した 9,155 事例すべてに手作業で正解ラベルを付与し[2]、それを用いて F 値（初期 seed F 値と区別するためにコーパス F 値と呼ぶことにする）を計算し、提案手法の性能を評価した。

表 3. 使用した素性セット

	feature sets									
	1	2	3	4	5	6	7	8	9	10
f1										
f2										
f3	○	○					○	○	○	○
f4										
f5			○		○		○			○
f6		○		○	○	○		○		○
f7						○		○	○	
f8	○	○	○	○	○	○	○	○	○	○
f9	○	○	○	○	○	○	○	○	○	○

[2] 部位記述表現は全体の 22.8%に相当する 2,085 事例あった。

素性セットについては表 2 に示す 9 種類すべての組み合わせを試すことはできないので、ステップワイズによって最適な 10 セットを用意した（表 3）。また、繰り返し回数は 5 回までとした。結果を表 4 に示す。また、図 2 に初期 seed F 値とコーパス F 値の閾値依存性を示す（図には 4 ラウンドまでを描いてある）。

表 4. 部位記述表現の抽出結果

round	feature set	seed Fmax	threshold	predicted F	corpus Fmax	threshold*
1	f3+f5+f8+f9	0.548	1,397	0.773	0.860	1,944
2	f3+f8+f9	0.483	1,742	0.851	**0.871**	1,985
3	f3+f5+f6+f8+f9	0.445	1,909	**0.858**	0.860	1,941
4	f5+f8+f9	0.365	2,356	0.786	0.802	2,038
5	f3+f8+f9	0.311	2,974	0.714	0.715	2,945

表 4 の feature set 列は各ラウンドで選択された最適な素性セットを示し、seed Fmax 列は初期 seed F 値の最大値、threshold 列はそのときの閾値を示す（図 2 の T_1, T_2, T_3, T_4）。predicted F 列は threshold 列の閾値に対応するコーパス F 値である。これは図 2 では T_1, T_2,…の破線とコーパス F 曲線の交点で与えられる。つまり、predicted F 列の値が提案手法によって達成できる性能である。

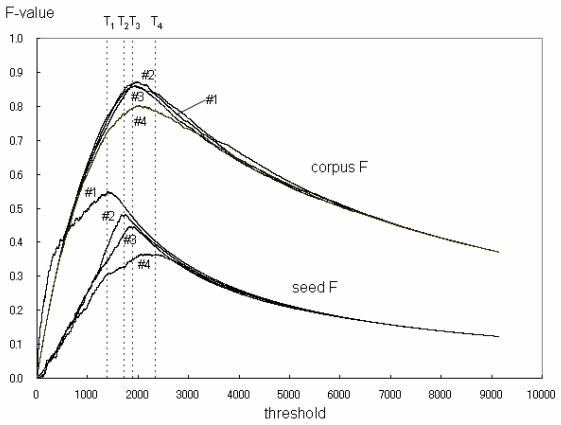

図 2. 初期 seed F 値とコーパス F 値の閾値依存性

さらに、表 4 の corpus Fmax 列には正解コーパスを用いて計算した実際の最大 F 値も示してある。これは図 2 におけるコーパス F 曲線の最大値であり、そのときの閾値の値が表 4 の threshold*列に示した数値である[3]。

これらの結果が示すように、提案手法はマーカである初期 seed の振る舞い（seed F 曲線）から実際の正解の振る舞い（corpus F 曲線）を推測していることがわかる。

表 4 より、提案手法は各ラウンドで f3+f5+f8+f9, f3+f8+f9, f3+f5+f6+f8+f9 の素性セットを使用して 3 回目のラウンドで最高性能 $F = 0.858$ に達していることがわかる。これは正解を知っていれば得られたであろう $F = 0.871$（表 4 の corpus Fmax の 2 ラウンド目の値）の 98.5%に相当する値である。

[3]用語インスタンス中の正例数にほぼ一致している。

得られた結果を比較評価するために同じ条件下で教師付き学習手法である最大エントロピーモデルを用いて分類実験を行った。結果を図 3 に示す。

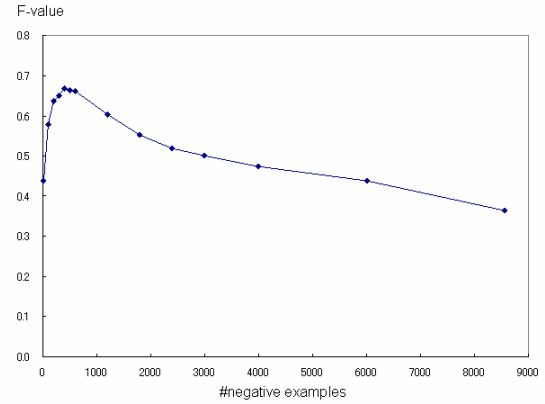

図 3. 最大エントロピーモデルによる分類性能

教師付き学習モデルは学習データ中に正例だけでなく負例を必要とする。正例は初期 seed を利用するとしても負例がない。そこで、分類対象の用語インスタンス集合から初期 seed を除いた残りから負例をランダムサンプリングしてモデルの学習を行い、学習データを含む用語インスタンス集合全体を対象として分類実験を行った[4]。図 3 の横軸はサンプリングした負例の数である。図に示すように負例数が 400 の場合に $F = 0.668$ で最大となった。これは提案手法の 8 割程度の値である。よって提案手法は同じ条件下では教師付き学習モデルよりも優れている。

4. 考察

Yarowsky[5]は文脈素性の冗長性を利用した教師なし学習モデルである決定リストの手法を WSD (Word Sense Disambiguation) 問題に適用して良好な成果を得た。Blum 等[6]は正解が付与されている例題が十分にない状況下であっても素性が持つ冗長性を利用すれば正解が付与されていない例題を用いて効率的に分類問題を解くことができる co-training という概念を理論的に定式化し、Web ページの分類問題でその有効性を示した。Collins 等[3]は Yarowsky の決定リストの手法を Blum 等の理論的な観点から定式化し、素性をスペル素性と文脈素性に分離して、それらを繰り返しの各ラウンドで交互に利用する DL-CoTrain アルゴリズムを提案した。彼らはそれを固有表現分類タスクに適用して Yarowsky のアルゴリズムを凌ぐ性能を報告している。

本研究で用いた手法もこれらの先行研究の延長線上にあり、大量の正解付き例題を必要としない教師なし学習モデルである。異なるのは、先行研究が少数の初期 seed しか必要としないのに対して、本研究では既存の用語リストを初期 seed として利用している点である。さらに重要な違いは、本研究では繰り返しの各ラウンドで最適な素性セットや閾値などモデルのパラメタをアルゴリズムが自動獲得している点である[5]。なぜこのようなことが可能になったかといえば、初期 seed をマーカとして正解データの挙動を推定しているからである。これに対して先に述べた先行研究では、各ラウンドにおいて決定リストに追加する学習によって獲得されたルールの数は数個に固定されており、その意味で閾値という概念がアルゴリズムには存在しない[6]。

2.4 節に示したアルゴリズムにおいて、初期 seed をマーカとして最適な素性セットを選択する際の基準として閾値最大基準を用いたことを述べた。これは、ラウンド数が小さい場合は seed F 曲線のピークがコーパス F 曲線のそれよりも小さい閾値に出現するため、高い性能を得るにはできるだけ大きな閾値を使った方が有利だからである（図 2 参照）。しかしながらラウンド数の増加に伴い seed F 曲線のピークは次第に右にずれ（$T_1<T_2<T_3<\ldots$）、いずれコーパス F 曲線のそれを追い越してしまい、返って性能は劣化する（図 2 の T_4）。これは各ラウンドで使用する学習データに含まれる負例数が次第に増加するためである。実際、ラウンド数の増加につれて負例数増加に伴うノイズの影響でコーパス F 曲線が次第に鈍っていくのが図 2 からもわかる。

本研究で提案したアルゴリズムの問題は最適な繰り返し回数を判定できないことである。つまり、アルゴリズムの終了条件が不明という問題である。繰り返しの終了を判定するためには初期 seed をマーカとして算出できる何らかの指標が必要となるが、これについては今後の課題として残された。

参考文献

[1] 田中昌昭，竹内孔一；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. 言語処理学会第 14 回年次大会発表論文集．pp.131-134 (2008).

[2] 田中昌昭，竹内孔一；自然言語で記述された画像読影レポートからの段階的局所化による固有表現抽出の試み. 第 28 回医療情報学連合大会論文集，pp.931-934 (2008).

[3] Collins M and Singer Y; Unsupervised Models for Named Entity Classification, Joint SIGDAT conference on EMNLP and VLC, pp.100-110, (1999).

[4] Torii M, Kamboj S and Vijay-Shanker K; Using name-internal and contextual features to classify biological terms, Journal of Biomedical Informatics 37, pp.498-511 (2004).

[5] Yarowsky D; Unsupervised Word Sense Disambiguation Rivaling Supervised Methods, Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics, Cambridge, MA, pp.189-196 (1995).

[6] Blum A and Mitchell T; Combining Labeled and Unlabeled Data with Co-Training, Proceedings of the 11th Annual Conference on Computational Learning Theory(COLT-98), pp.92-100 (1998).

[4] 通常の教師付き学習手法の評価を行う場合は検証用データに学習データを含めないが、提案手法と条件を同じにするためにこのようにした。

[5] 各ラウンドで使用する素性セットを変えることによって性能が向上する理由は解明できていない。実験した範囲内では素性セットを固定した場合、今回得られた性能を上回ることはなかった。

[6] これは閾値が存在しないのではなく暗黙の自明な閾値（追加するルールの個数）を設定しているとも考えられる。