人手の規則構築による専門用語抽出の精度比較

山崎 元揮　　竹内 孔一　　源 翔三郎
岡山大学大学院 自然科学研究科
{yamasaki, koichi, minamoto}@cl.cs.okayama-u.ac.jp

1　はじめに

本稿は規則ベースモデル [1] による専門用語抽出を行い，統計的学習モデル [2] と精度比較し，結果規則ベースモデルは専門用語に関連が高い語において再現率が高い事を明らかにする．

用語抽出の方法は規則ベースモデルと統計的学習モデルがある．統計的学習モデルを利用した用語抽出システムでは，SVM[3] や CRF[4] など近年発展した機械学習の枠組みを利用する事ができる．しかし，学習データを用意する必要があり学習データに存在しない分野の用語を抽出する事はできない．規則ベースモデルを利用した用語抽出システムでは，人手によって用語出現パターンを作成する為，学習データに存在しない用語を抽出する事ができる．しかし，人手で規則を構築する為時間がかかり，人によって精度が変わる為安定しない．そこで，(1) 人手による規則ベースの用語抽出を 2 人の別の作業者が構築したものを比較 (2) 規則ベースと統計的学習モデルの用語抽出の精度の比較，を行う．

規則ベースによる手法では，先行研究 (新納 2010) が SRL 上で構築した用語抽出してステムと，著者が SRL を利用して，先行研究を参照せずに同じ開発コーパスを利用して構築したシステムを構築する．規則作成の期間はどちらも 3 カ月程度で行い，規則の組み方による精度の違いの確認を行った．

2　用語抽出手法

この章では，まず本研究で使用する規則ベースモデルの SRL と統計的学習モデルの CRF を利用した用語抽出について述べる．SRL では規則の組み方，抽出の仕組みを説明し，CRF ではテンプレートや使用するツールについての説明を行う．

2.1　SRL

SRL はテキストから規則に基づいて用語を抽出する為の言語であり，抽出した用語がどのクラスに属するかについて指定できる．SRL では用語抽出のために Entity Rules と Template Rules という 2 種類の規則を使用する．

Entity Rules は語構成を規則で記述し，テキスト中の単語列に対してクラス分類する．語構成の規則は単なる表層を利用するだけでなく，文字の種類や部分一致など関数で用意されており利用できる．更に，意味のある単語集合，例えば接尾辞集合など自由にリスト化して，それを 1 つの関数クラスとしてパターンの中に埋め込んで記述できる．以下に具体例を挙げて説明する．

例えば「レジオネラ菌」の様にカタカナ語と感染症に関連性の高い「菌」から構成される用語を取り上げる．この語は BACTERIA クラスに分類させたいので，Entity Rules は下記の様になる．

```
R1 :- name(bacteria,X){ ortho(”InKatakana”) ”菌” }
```

R1 は規則の番号であり人手による記述部は :- 以降である．まず最初の name(bacteria,X) であるが，これは「X という単語列は BACTERIA クラスに分類される」という宣言である．この bacteria という部分を他のクラスに換えることで，それぞれのクラスに応じた宣言となる．

続く中括弧内では単語列 X の内容を記述する．上の例では ortho(”InKatakana”) と「菌」という関数と文字が記述されている．この ortho は規則の記述のために SRL が用意している関数で，ortho(”InKatakana”) は単語の文字列の種類を判定を行い，全てカタカナ文字列で構成された単語の場合マッチする (以降カタカナ語で構成された用語を【カタカナ】と表す)[1]．この ortho(”InKatakana”) と後ろに記述されている「菌」という文字を合わせて中括弧内は「【カタカナ】 - 菌

[1] 他にも漢字，英字，ひらがなで構成されたものはそれぞれ【漢字】,【英字】【ひらがな】と表す.

」となり,「単語列 X は【カタカナ】と「菌」という文字が連続したもの」ということを表している．

この Entity Rule によりテキスト中に「【カタカナ】- 機能 」という構成の単語列である「カンピロバクター菌」や「コレラ菌」が現れた際に，それらの単語列を BACTERIA クラスに分類することができる．

Entity Rule で利用する重要な関数として単語リストがある．単語リストはある概念を表す中心的な語のグループである．例えば「菌」という概念であれば @bacteria[2] として登録することで先ほどの規則を下記の様に．書き換える事ができる．

R1 :- name(bacteria,X){ ortho(”InKatakana”) list(@bacteria) }

この時，@bacteria に「球菌」を登録すれば「ブドウ球菌」等の語も抽出可能となり規則をまとめる事が可能となる．

Template Rules は Entity Rules での分類結果を要素として扱い，文脈を規則で記述して用語を抽出する．例えば，先の Entity Rule の結果で BACTERIA クラスに分類にされた「レジオネラ菌」を BACTERIA クラスに属する用語として抽出するための規則は次のように記述できる．

T1 : bacteria(X) :- name(bacteria,X)

name(bacteria,X) が Entity Rules の結果を要素としている部分である．Entity Rules がクラス分類を行った用語を，この Template Rules によって実際に抽出する事となる．

テキストから用語抽出までの全体処理の流れを図 1 に示す．

2.2 CRF による用語抽出システム

CRF とは複数の素性をもとに時系列ラベルを柔軟に予測する確率モデルでデータスパースネスに強く，形態素解析等で良い精度が示されている [5]．CRF を利用するツールとしてオープンソースの CRF ツールキットである CRF++[3] を利用し，クラスの抽出規則を学習して用語の抽出を行う．CRF では学習モデル作成の際にテンプレートファイルが必要となる．テンプレートファイルとは作成したコーパスの中からどの特徴（要素）を学習とテストで使用するかを記述したものである．本研究で使用する特徴量については 3.1 節で述べる．

[2]単語リストには先頭に@を付ける
[3]http://crfpp.sourceforge.net/.

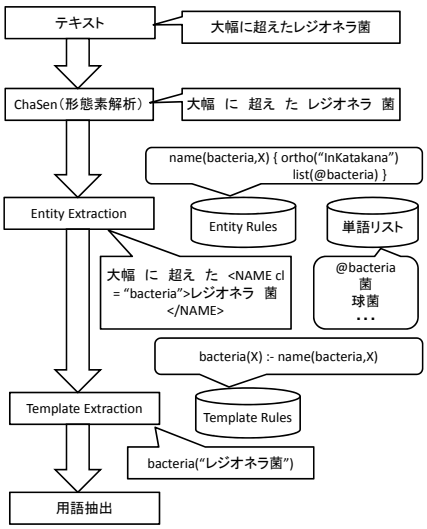

図 1: 用語抽出までの処理手順

3 実験

この章では SRL を利用した感染症用語抽出に関する実験を行い，精度評価を行う．まず用語抽出の対象データについて述べ，SRL の規則構築の方針について説明する．最後に実験結果から得られた考察を述べる．

3.1 実験準備

今回実験に用いた対象データは専門家が感染症用語の抽出及びクラス分類を行い，タグ付与されたニュース記事 [6] である．学習用データ（CRF）及び開発用データ (SRL) として BioCasterj_1003(100 件)，BioCasterj_1102(200 件)，WHO parallel corpus_ja_1213(50 件) を合わせた合計 350 件の記事を使用した．テスト用データとして BioCasterJ_0731(100 件)，BioCasterJ_1018(98 件) を合わせた合計 198 件の記事を使用した．

精度比較では CRF と SRL，及び先行研究との比較を行った．CRF のテンプレートにおいて使用した特徴量は基本形，品詞，品詞のサブタイプである．規則ベースの SRL では，感染症分野に対して先行研究の構築した SRL-S と著者が構築した SRL-Y を比較する．SRL-S では用語の部分構造を修飾部と主部にわけて組合せによる規則を構築した．さらに Web 上の感染症に関する用語リストを利用して網羅性を挙げた．一方，Y モデルでは末尾の形態素が所属する単語リストと文字種による組合せで規則を構築した．例えば「鳥インフルエンザ」という用語があった場合,「鳥」は【漢字】であり,「インフルエンザ」は@disease に登録された形態素である．つまり「鳥インフルエンザ」に対して

「【漢字】 - @disease」の組合せで規則を作成する．比較の枠組みとしてタグ付与された全てのクラス，感染症分野と特に関連性の低いクラスを除いた特定 14 クラス，関連性の高いものだけを残した特定 10 クラスで行った．一方，人工知能分野における規則も著者は構築した．こちらは著者が構築した SRL と CRF で比較を行った．人工知能分野のコーパスでは用語のクラス分類がされてない為 [7]，クラスにおける比較は無い．

3.2 結果

表 1 に CRF と SRL-Y の全クラスにおける用語抽出精度を示す．適合率では DNA と RNA 以外で CRF が勝っているが，再現率では SRL が勝っているクラスの方が多い．しかし，全体平均で見ると適合率，再現率共に CRF が勝り F 値でも CRF が 0.1 程高い値を獲得している．

表 1: CRF と SRL の全クラスにおける用語抽出精度

クラス	適合率 (CRF)	適合率 (SRL)	再現率 (CRF)	再現率 (SRL)	F 値 (CRF)	F 値 (SRL)
ANATOMY	0.8780	0.6225	0.1915	0.5054	0.3144	0.5579
BACTERIA	0.8587	0.6429	0.6475	0.8852	0.7383	0.7448
CHEMICAL	0.9000	0.6486	0.0849	0.4528	0.1552	0.5333
CONDITION	0.8621	0.7786	0.7194	0.5072	0.7843	0.6143
CONTROL	0.8780	0.5217	0.2647	0.3529	0.4068	0.4211
DISEASE	0.7885	0.7396	0.5672	0.5949	0.6598	0.6594
DNA	0.0000	0.5000	0.0000	0.2500	0.0000	0.3333
LOCATION	0.9207	0.7366	0.6609	0.5440	0.7695	0.6528
NON_HUMAN	0.9091	0.5722	0.3672	0.5751	0.5185	0.5736
ORGANIZATION	0.8016	0.6372	0.5231	0.4346	0.6330	0.5168
OUTBREAK	0.8353	0.7907	0.5635	0.5397	0.6730	0.6415
PERSON	0.7481	0.5870	0.5885	0.4644	0.6588	0.5168
PRODUCT	0.5854	0.5500	0.4444	0.6111	0.5053	0.5789
PROTEIN	1.0000	0.5152	0.0541	0.4595	0.1026	0.4857
RNA	1.0000	1.0000	0.5000	1.0000	0.6667	1.0000
SYMPTOM	0.8714	0.4982	0.5755	0.6651	0.6932	0.5697
TIME	0.8395	0.5606	0.7826	0.6330	0.8101	0.5946
VIRUS	0.7700	0.6667	0.5923	0.5077	0.6696	0.5764
全体平均	0.8205	0.6298	0.5816	0.5296	0.6807	0.5754

表 2 に CRF と SRL-Y の特定 14 クラスにおける用語抽出精度を示す．適合率，再現率は特定 10 クラスのものと同様に適合率では CRF，再現率では SRL が良い結果が出ている．全体平均の F 値ではほぼ同様の値が出ているが 0.01 程 SRL が低い値となった．

表 3 に CRF と SRL の特定 10 クラスにおける用語抽出精度を示す．適合率では PRODUCT と RNA 以外は CRF が上回っているが，再現率では VIRUS 以外 SRL が勝っている．全体平均の F 値においても SRL が 0.05 程高い値を獲得している．

表 4 と表 5 に [1] の感染症用語抽出の結果を示す．適合率では本手法が上回っているが，再現率では [1] が大きく上回っており結果 F 値は新納（2010）が高い値を獲得している．

表 2: 特定 14 クラスにおける用語抽出精度

クラス	適合率 (CRF)	適合率 (SRL)	再現率 (CRF)	再現率 (SRL)	F 値 (CRF)	F 値 (SRL)
ANATOMY	0.8824	0.6225	0.1596	0.5054	0.2703	0.5579
BACTERIA	0.8571	0.6429	0.5902	0.8852	0.6990	0.7448
CHEMICAL	0.8750	0.6316	0.0660	0.4528	0.1228	0.5275
CONDITION	0.8630	0.7107	0.7098	0.5435	0.7789	0.6158
CONTROL	0.8718	0.5161	0.2500	0.3529	0.3886	0.4192
DISEASE	0.8138	0.6645	0.5356	0.5949	0.6460	0.6277
DNA	0.0000	0.5000	0.0000	0.2500	0.0000	0.3333
NON_HUMAN	0.9167	0.5635	0.3420	0.5751	0.4981	0.5692
OUTBREAK	0.8353	0.7907	0.5635	0.5397	0.6730	0.6415
PRODUCT	0.5909	0.5410	0.2407	0.6111	0.3421	0.5739
PROTEIN	1.0000	0.5152	0.0270	0.4595	0.0526	0.4857
RNA	1.0000	1.0000	0.5000	1.0000	0.6667	1.0000
SYMPTOM	0.8657	0.4982	0.5472	0.6651	0.6705	0.5885
VIRUS	0.7629	0.6667	0.5692	0.5077	0.6520	0.5764
全体平均	0.8396	0.6248	0.4711	0.5668	0.6035	0.5944

表 3: 特定 10 クラスにおける用語抽出精度

クラス	適合率 (CRF)	適合率 (SRL)	再現率 (CRF)	再現率 (SRL)	F 値 (CRF)	F 値 (SRL)
ANATOMY	0.8146	0.6225	0.1489	0.5054	0.2545	0.5579
BACTERIA	0.8750	0.6429	0.5984	0.8852	0.6986	0.7448
CHEMICAL	0.8391	0.6154	0.0660	0.4528	0.1228	0.5217
DISEASE	0.8750	0.6544	0.4881	0.5988	0.6099	0.6254
DNA	0.8125	0.5000	0.0000	0.2500	0.0000	0.3333
PRODUCT	0.0000	0.5323	0.2222	0.6111	0.3200	0.5690
PROTEIN	0.5714	0.5152	0.0270	0.4595	0.0526	0.4857
RNA	1.0000	1.0000	0.5000	1.0000	0.6667	1.0000
SYMPTOM	1.0000	0.4965	0.5472	0.6651	0.6705	0.5685
VIRUS	0.7526	0.6600	0.5615	0.5077	0.6432	0.5739
全体平均	0.7526	0.6054	0.4100	0.5982	0.5455	0.6014

表 4: 新納（2010）による感染症用語抽出精度（全クラスと特定 14 クラス）

クラス	適合率 (全)	適合率 (14)	再現率 (全)	再現率 (14)	F 値 (全)	F 値 (14)
ANATOMY	0.7169	0.6515	0.8441	0.8441	0.7753	0.7354
BACTERIA	0.5125	0.5093	0.6721	0.6721	0.5816	0.5795
CHEMICAL	0.6771	0.6600	0.6132	0.6226	0.6436	0.6408
CONDITION	0.6107	0.5324	0.9351	0.9471	0.7388	0.6817
CONTROL	0.2531	0.2450	0.4485	0.4485	0.3236	0.3169
DISEASE	0.7231	0.6645	0.8103	0.8142	0.7642	0.7318
DNA	0.4286	0.4286	0.7500	0.7500	0.5455	0.5455
LOCATION	0.6760		0.7187		0.6967	
NON_HUMAN	0.6359	0.6188	0.7150	0.7150	0.6732	0.6635
ORGANIZATION	0.4958		0.6051		0.5450	
OUTBREAK	0.7833	0.7787	0.7460	0.7540	0.7642	0.7661
PERSON	0.5373		0.5998		0.5668	
PRODUCT	0.5122	0.5060	0.7778	0.7748	0.6176	0.6131
PROTEIN	0.6304	0.6304	0.7838	0.7838	0.6988	0.6988
RNA	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
SYMPTOM	0.5495	0.5495	0.7075	0.7075	0.6186	0.6186
TIME	0.6555		0.7494		0.6993	
VIRUS	0.5460	0.5398	0.7308	0.7308	0.6250	0.6209
全体平均	0.5908	0.5672	0.7006	0.7740	0.6410	0.6547

表 5: 新納（2010）による感染症用語抽出精度（特定 10 クラス）

クラス	適合率 (10)	再現率 (10)	F 値 (10)
ANATOMY	0.6255	0.8441	0.7185
BACTERIA	0.4970	0.6721	0.5714
CHEMICAL	0.6600	0.6226	0.6408
DISEASE	0.6478	0.8142	0.7215
DNA	0.4286	0.7500	0.5455
PRODUCT	0.5000	0.7778	0.6087
PROTEIN	0.6304	0.7838	0.6988
RNA	1.0000	1.0000	1.000
SYMPTOM	0.5474	0.7075	0.6173
VIRUS	0.5398	0.7308	0.6209
全体平均	0.5962	0.7638	0.6697

表 6 に人工知能分野における SRL-Y と CRF の用語抽出精度の結果を示す．適合率，再現率共に CRF が上回る結果となった．

表 6: 人工知能分野における SRL と CRF の用語抽出精度

	適合率	再現率	F 値
SRL	0.7081	0.6753	0.6913
CRF	0.8719	0.8322	0.8516

3.3 考察

実験結果から SRL は CRF に比べ再現率が良く，専門的な用語であれば CRF よりも良い精度を獲得している事が分かる．例えば「眼科薬」や「緑内障治療薬」の様な語は学習データに無いため，CRF では用語抽出できなかったが SRL では抽出できた．これは「【漢字】 - @chemical」や「【漢字】 - 【漢字】 - @chemical」の規則によって抽出が可能となっており，パターン抽出の利点といえる．

しかし,【ひらがな】を挟まずに文字列が続いた場合に SRL は抽出ミスが頻発する．例えば「宇陀市榛原区」の様な語は「宇陀市」と「榛原区」でそれぞれ用語として抽出すべき語であるが，SRL では「宇陀市榛原区」を１つの用語として抽出してしまう．これは SRL が最長パターンマッチを行い,「宇陀市」と「榛原区」で分けて抽出する規則よりも「宇陀市榛原区」と抽出する規則を優先する為である．

[1] との比較では本手法が適合率が高く，再現率は [1] が良い結果となった．これは [1] が見つかった用語ほぼ全てに対応する規則を作成したのに対して，本研究では用語以外のものを取らないような規則構築を行った為と考えられる．F 値で見ると新納の方が値が良い事から，SRL の長所である再現率を伸ばす様に規則を構築するのが良いと考えられる．

4 おわりに

本稿では規則ベースモデルの SRL を用いて感染症用語抽出システムを構築した．用語抽出規則の作成方針は単語リストと文字種の組合せで行った．この規則を用いて得られた精度を CRF のものと比較を行った結果，SRL は関連性の高いクラスにおける再現率で CRF よりも高い値を獲得した．また先行研究である新納（2010）との比較も行い，規則の組み方により精度にも差が出る事を明らかにした．

参考文献

[1] 新納貴志, 規則に基づく感染症用語抽出システムの構築, 岡山大学工学部情報工学科特別研究報告, 2010.

[2] 岡田和也, 感染症情報提示システムに向けた記事分類と感染症用語抽出, 岡山大学大学院自然科学研究科電子情報システム工学専攻修士論文, 2008.

[3] 甘利俊一, 麻生英樹, 津田宏治, 村田昇, 統計科学のフロンティア 6, 第 6 巻, pp.107-120, 岩波書店, 第 10 版, September 2001.

[4] Pereira. F Lafferty. J, McCallum. A. Conditional random elds: Probabilistic models for segmenting and labeling sequence data, 2001.

[5] 鹿島久嗣, 坪井裕太, 工藤拓, 言語処理における識別モデルの発展-HMM から CRF まで-. 言語処理学会第 12 回年次大会チュートリアル, 2006

[6] A. Kawazoe and L. Jin and M. Shigematsu and R. Barrero and K. Taniguchi and N. Collier. The development of a schema for the annotation of terms in the BioCaster disease detection/tracking system. Proceedings of the International Workshop on Biomedical Ontology in Action (KR-MED 2006), pp.77-85, 2006

[7] Kageura, K. and Yoshioka, M. and Takeuchi, K. and Koyama, T. Overview of the TMREC Tasks. *Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition, pp.415-415, August 30 - September 1, KKR Hotel Tokyo, Tokyo, Japan, 1999*