統計的手法を利用した伝染病検索システムの構築に向けて

竹内 孔一，岡田和也　　川添愛†，コリアー・ナイジェル†
岡山大学大学院自然科学研究科　†国立情報学研究所
koichi@it.okayama-u.ac.jp　　collier@nii.ac.jp

1　はじめに

本研究報告では，統計的手法を利用したニュース記事からの伝染病に関連した固有表現抽出について報告する．これは現在構築中の多言語伝染病情報提示システム (BioCaster システム) 構築の一環で行っている研究である．伝染病はその拡大が心配されるため早く正確にその状況を集約する必要があるが，そうした事件を一番に知る手がかりとして Web 上のローカルニュースの活用が考えられる．我々は英語，日本語，タイ語，ベトナム語というアジア圏の多言語ニュースサイトから自然言語処理技術を利用して伝染病情報を集約し伝染病の監視および対応を行う専門家に流行の可能性をいち早く伝える情報提示システムの開発を行っている．このシステムを構築するために，伝染病のニュースから誰がどんな病気にかかり，どうなってるかという事態を抽出する必要がある．しかし事態はさまざまな表現の多様性がありそれらを集約するのは容易ではない．

そこで我々はこの事態抽出を行う前段階として病名，症例，ウィルス，感染者，薬物といった必要な要素の体系化を行い，それに基づく正しい正解付きデータを作成することで，従来研究されてきている固有表現抽出法に応用し，事態に必要な要素をまず同定することからはじめている．

結果として 500 記事で学習させた CRF モデルを 50 記事の WHO(World health organization) のリポート記事に適用したところ約 70%程度の正確さで上記の事態要素を認識することに成功した[1]．

本論文ではどのように伝染病関連の事態要素を定義したか，どのように事態要素を抽出するシステムを作成しどう評価したかについて以下に記述する．

2　伝染病関連の固有表現体系の整理

伝染病関連ニュースから事態を抽出するために，要素 (固有表現) をカテゴリー化して，それらを具体的にニュース記事に対して意味タグとして付与する必要がある．ここで問題となるのは

- 単に固有表現を抽出するだけではなく事態として後に集約する必要がある
- 意味タグをゆれを少なくテキストデータに付与し正解データを作成する

という 2 点である．前者に対してはただ単に意味タグを設定するだけでなく，意味タグが伝染病という事態に対してどういう役割関係を担っているかをまとめた体系を作成する．後者に対しては，各意味タグの事例と定義について英語で作成し [3]，多言語におけるタグ付与のゆれを少なくすることを目指している．

まず図 1 に意味タグとその体系を示す．図の中で大

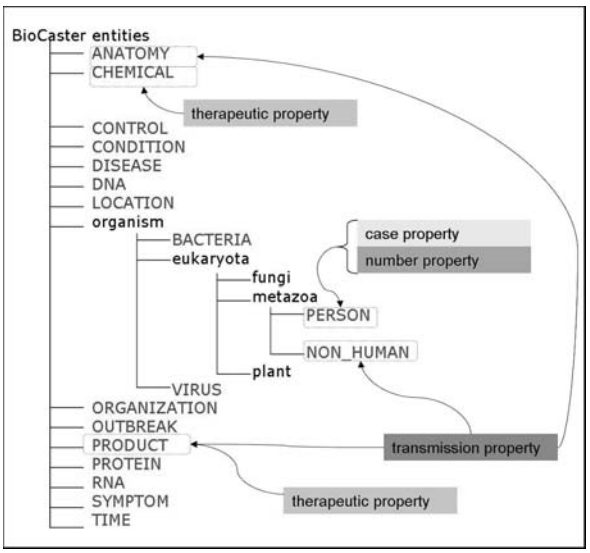

図 1: 意味タグの体系

文字で記述された葉の部分が意味タグである．18 種類あり，これらは ANATOMY, BACTERIA, CHEMICAL, CONDITION, CONTROL, DISEASE, DNA, LOCATION, NON_HUMAN, ORGANIZATION, OUTBREAK, PERSON, PRODUCT, PROTEIN, RNA, SYMPTION, TIME, VIRUS である．これらの意味タグについて表 1 にその事例と説明を示す (詳細は Kawazoe et al.[3] を参照)．

表 1 に示す意味タグを具体的にテキストに統一的に付与することが重要である．そのため，基本的な意味タグの考え方と詳細なタグ付与事例を記述したマニュ

[1]同様の結果を英語，タイ語，ベトナム語で WHO の記事に対して評価を行っている．結果は現在投稿中である．

表 1: 意味タグの説明

Class	Example	Description
ANATOMY	患者の [肝臓]	身体の一部
BACTERIA	調理師からの [コレラ菌]	感染に関与のバクテリア
CHEMICAL	[抗生物質] に代わる	化学物質の一部
CONDITION	2500 人が [重症]	患者の病状
CONTROL	効果的な [ワクチン接種]	感染を弱める手段
DNA	[vanA 遺伝子]	遺伝子情報
DISEASE	ひどい [肺炎]	感染の病気
LOCATION	[川崎市] の	場所
NON_HUMAN	[家禽]7 頭の	感染に関与の動物や菌
ORGANIZATION	[厚生労働省] は	組織
OUTBREAK	米国で [流行] 中	爆発的な感染を示す表現
PERSON	同区内の [客 5 人]	感染者や集団
PRODUCT	効果的で安全な [ワクチン]	生物製剤
PROTEIN	不可欠な [筋肉タンパク]	たんぱく質
RNA	1 本の [RNA]	核酸の情報
SYMPTOM	激しい [せき]	症状
VIRUS	臨床材料の [H5N1]	感染のウィルス
TIME	日本時間の [11 日朝]	時間

アルを作成している [3]．例えば PERSON というタグはここでは氏名，性別や職業，人数，年齢を付与対象にする一方で，修飾詞は対象としない．例を示すと「[インドネシア]LOCATION 旅行中の [18 歳から 42 歳の男性 5 人]PERSON が」となる．

このように本研究で定義している意味タグは従来の固有表現抽出タスクで見られるような単なる名前ではなくより複雑な単位であることがわかる．

以下の節では上記の意味タグを付与したデータを利用してどのように固有表現抽出モデルを構築し，付与実験を行ったかについて述べる．

3 固有表現の抽出

前節で設定した意味タグをテキストデータに付与するために固有表現抽出タスクで用いられてきた統計的学習モデルを利用した枠組みを利用する．つまり，人手で作成した学習用データを作成し，統計的学習モデルを利用して新しいニュースから伝染病関連の情報を取り出す．新しいニュースに対して我々が設定した意味タグがどの程度精度良く予測できるかを評価するために分野のバランスを取った新聞記事コーパス以外に WHO(世界保健機構) の記事コーパスを用意した．以下学習モデル，特徴量，コーパスについて記述し実験結果と考察について述べる．

3.1 抽出モデル

SVM と CRF は入力される各単語に対して目標とする意味タグを付与することで固有表現をテキストから抽出する．意味タグの付与は分類しようとする単語の文脈情報をコード化した特徴量を基に行われる．特徴量は固有表現抽出では通常，表層の単語や品詞など膨大な組み合わせが可能となるが SVM などのカテゴリー分類器では HMM などの生成的モデルと比較して膨大な組み合わせの特徴量に対して良い精度を得られることが示されている [5]．これにより自由に特徴量を設定することができる．以下 SVM と CRF について簡単に説明する．

SVM はカーネル法を利用して高次元空間で判別を行うことで高い識別能力を示す．SVM の判別式は $f(x) = sign(\sum_{i=1}^{N} \alpha_i y_i k(x_i, x_j) + b)$ で示され，N 個の学習データ (x_i, y_i が入力，出力) で学習した結果最適な α と b が決定される．我々が先に行った分子生物学文献における固有表現抽出の実験結果 [2] から，高速で高い精度を示す多項式カーネル $k(x_i, x_j) = (x_i \dot{x}_j + 1)^d$ (d = 2) を利用する．パッケージとして TinySVM [2]を利用した．

一方，CRF は HMM(隠れマルコフモデル) をベースに多数の特徴量を扱いつつ学習データにおけるデータスパースネスを巧妙に回避したカテゴリー分類器である．SVM との大きな違いは Viterbi アルゴリズムの利用によってある意味タグを決定するために全体の意味タグの組み合わせを動的に考慮して決定できることである．CRF では入力 x に対する出力 y の評価式は以下のようになる．

$$P(y|x) = \frac{exp\langle \Theta, \Phi(x,y) \rangle}{\sum_{y \in Y} exp\langle \Theta, \Phi(x,y) \rangle} \quad (1)$$

$$\hat{y} = \underset{y \in Y}{argmax}\, P(y|x)$$

この式に対する詳細は文献 [4] に譲る．パッケージとしては CRF++を用いた[3]．

次に特徴量について述べる．SVM と CRF の分類能力を比較するために，同じ特徴量を利用する．特徴量としては，表層の単語，活用の基本形，品詞，係り先の情報，文字種情報，場所情報，ニュース元，ニュースの種類である．分かち書き，基本形，品詞は形態素

[2]TinySVM (http://chasen.org/~taku/software/TinySVM/)
[3]CRF++(http://chasen.org/~taku/software/CRF++/)

表 2: balanced コーパスの内容

分野	記事数
経済	50
健康	249 (244)
娯楽	0
政治	50
スポーツ	50
科学技術	50
社会	51
合計	500 (244)

解析システム ChaSen を利用して取り出し，係り先の情報と場所情報は係り受け解析システム CaboCha[4] が出力する情報を利用した．

3.2 データ

伝染病関連のニュースはさまざまな分野のニュースに現れるため，一般記事の中から正確に伝染病に関係する部分を取り出す必要がある．そこで上述の伝染病関連の固有表現を抽出するシステムを評価する学習データとして半分程度伝染病の記事を含まない他の分野をまぜた記事コーパス (以下，balanced コーパス) を作成した (表 2 参照)．ここで () 内は直接伝染病関連の記事である．これらは一般の新聞記事ニュースサイトから取得した．さらに新聞以外の記事として別に WHO(世界保健機構) の記事 50 記事を評価用コーパスとして用意し，新聞記事と異なる伝染病関連のニューステキストに対してどの程度固有表現抽出モデルが有効であるか評価を行うために利用する．以下ではこれらの記事に対する意味タグの統計量を表 3 に示す．意味タグでは Location が最も多く DNA や RNA などの専門性の高い用語は数が少ない．こうした統計量のばらつきがある中でどの程度学習が成功したかについて抽出実験による評価を次節で行う．

3.3 抽出結果と考察

前節で用意した 2 つのコーパス (balanced コーパス 500 記事と WHO コーパス 50 記事) に対して SVM と CRF を利用して学習と評価を行う．分析視点として，(a) 特徴量の有効性，(b) 統計モデルの違いによる精度の異なり，(c) コーパスの違いによる精度の異なりについて調べた．(a)(b) は balanced コーパスに対して 10 回の交差検定で精度を求めた．(c) は balanced コーパスで得られた最も良い特徴量の組み合わせで balanced コーパスを全てを学習したモデルを WHO コーパスに適用することでコーパスの質の違いによる精度の異なりを評価する．精度は適合率，再現率，F 値で評価する．

[4]CaboCha(http://chasen.org/ taku/software/cabocha/)

表 3: コーパス内のタグ出現回数

Class	Balanced News	WHO
ANATOMY	480	11
BACTERIA	326	9
CHEMICAL	237	3
CONDITION	891	158
CONTROL	260	42
DISEASE	1216	97
DNA	16	0
LOCATION	2224	429
NON_HUMAN	460	60
ORGANIZATION	2142	225
OUTBREAK	244	44
PERSON	3535	444
PRODUCT	125	31
PROTEIN	73	1
RNA	5	0
SYMPTON	526	25
TIME	1963	209
VIRUS	305	99
Total	15023	1886

まず特徴量を変えた場合の SVM と CRF による抽出実験の結果を表 4 と表 5 にそれぞれ示す．特徴量で「単，品，場，字，主，元，尾」はそれぞれ，単語，品詞，場所情報，文字種，主辞 (係り先)，記事元，特殊な接尾辞 (病名など) である．表 4 より SVM の場合は単語，品詞，場所，文字種を用いた場合が最も良く，特に品詞が有効であることが示されている．一方，CRF の場合も表 5 より品詞を利用した場合に精度が大きく向上している．このことから，品詞はかなり有効である．これは形態素解析 ChaSen の品詞の種類の豊富さと精度が関連しており，固有名詞では場所や人名に関してかなり網羅している部分が効果を発揮していると推測できる．ただし単語と文字種を組み合わせた場合精度の向上が見られているにもかかわらず，CRF の場合，品詞，場所に文字種を加えると精度が少し下がっている．特徴量の組み合わせによる精度への影響はもう少しさまざまな組み合わせを調べる必要がある．

同様に表 4，表 5 よりモデルの比較を行うと CRF

表 4: balanced コーパスに対する SVM の結果

	SVM		
特徴量	適合率	再現率	F 値
単	74.46	58.92	65.78
単, 字	74.50	60.32	66.67
単, 品	73.79	66.44	69.93
単, 品, 場	75.20	68.23	71.55
単, 品, 場, 字	75.36	**68.50**	**71.77**
単, 主	**77.54**	57.44	65.99
単, 元	73.53	49.46	59.14
単, 尾	73.28	60.41	66.22

表 5: blanced コーパスに対する CRF の結果

モデル	CRF		
特徴量	適合率	再現率	F 値
単語	80.18	64.46	71.47
単, 字	79.56	66.32	72.34
単, 品	79.90	70.82	75.09
単, 品, 場	**80.38**	**71.52**	**75.69**
単, 品, 場, 字	79.93	71.51	75.49
単, 主	80.08	66.41	72.51
単, 元	75.11	62.28	68.10
単, 尾	80.07	64.58	71.50

表 6: WHO コーパスに対する精度

モデル	適合率	再現率	F 値
SVM	66.70	61.84	64.18
CRF	76.61	68.39	72.27

の結果が SVM に比べて F 値で約 4%から 5%程度精度が高いことがわかる．内容として適合率，再現率ともに向上しての結果であることから，このタスクには CRF が有効であることがわかる．どちらのモデルとも適合率が単語のみの特徴量の場合でも精度が高く特徴量を増やすことで再現率が向上することで F 値全体が向上している．

次にコーパスの異なりによる意味タグ付与の結果を表 6 に示す．表から CRF が SVM よりも約 8%近く高い精度で WHO コーパスの意味タグを付与できたことを示している．balanced コーパスでの精度よりは劣っているが，SVM の場合は精度の下がり方が大きい．これより CRF モデルの特徴である全意味タグ列の組み合わせを考慮して解く方法が学習データに対して過度にパラメータが調整されることがなく，少し質の違うテキストデータに対する適応能力が高いことが推測できる．この結果からも人手によるタグ付きコーパスを基に未知のニュースに対して伝染病情報を抽出するシステムの構築に CRF が有効であることがわかる．

4 まとめ

伝染病情報を Web 上のニュースサイトの記事から抽出し集約するためのシステムを構築する第一段階として伝染病情報に必要な固有表現抽出システムを提案した．固有表現集合の定義を示し，SVM と CRF を利用した付与モデルをニュースサイトの記事に対して適用したところ CRF を利用したモデルが SVM に比べて高い精度を示すことを明らかにした．この結果は英語，ベトナム語，タイ語に対して同様に行われており [1]，今後多言語の観点による比較を行いたい．

参考文献

[1] Collier, N., Kawazoe, A., Jin, L., Shigematsu, M., Dien, D., Barrero, R., Takeuchi, K. and Kawtrakul, A.: A multilingual ontology for infectious disease surveillance: retionale, design and challenges, *Language Resources and Evaluation* (accepted to appear).

[2] Collier, N. and Takeuchi, K.: Comparison of character-level and part of speech features for name recognition in biomedical texts, *Journal of Biomedical Informatics*, No. 37, pp. 423–435 (2004).

[3] Kawazoe, A., Jin, L., Shigematsu, M., Barerro, R., Taniguchi, K. and Collier, N.: 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).

[4] Lafferty, J., McCallum, A. and Pereira, F.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, *Proc. 18th International Conf. on Machine Learning*, pp. 282–289 (2001).

[5] Mayfield, J., McNamee, P. and Piatko, C.: Named entity recognition using hundreds of thousands of features, *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4*, pp. 184–187 (2003).