自動構築した大規模訓練データを用いた固有名抽出

宇佐美 佑　　Han-Cheol Cho†　　岡崎 直観†　　辻井 潤一†
東京大学 理学部情報科学科　　東京大学大学院 情報理工学系研究科コンピュータ科学専攻
{yusmi, hccho, okazaki, tsujii}@is.s.u-tokyo.ac.jp

1　はじめに

固有表現抽出（NER）は、文書中で言及される実体・概念に対して意味クラスを付与するタスクであり、質問応答や情報抽出などのアプリケーションにおいて、基盤技術となっている。近年では、テキスト中で実体・概念の出現箇所を付与した訓練データを用意し、サポートベクトルマシン（SVM）や条件付き確率場（CRF）などの機械学習アルゴリズムに基づいて固有表現抽出器を構築するのが一般的である。また、IREX[1]、CoNLL 2003[2]、GENIA[3]、OntoNotes[4]に代表されるコーパスが整備されたことにより、人名、地名、組織名、遺伝子名など、特定の意味クラスの固有表現抽出器を、容易に構築できるようになった。

しかしながら、現状の訓練データの整備は、限られたドメインと意味クラスに限定されている。機械学習は、固有表現抽出器をドメインに依存することなく設計できるが、抽出したい意味クラス・ドメインのタグ付きコーパスを準備する必要がある。今後、情報抽出の応用範囲を様々なドメイン・意味クラスに拡張する際、訓練データの入手性が固有表現抽出器のボトルネックとなると考えている。

一方で、実体・概念の表現事例を収録している語彙データベースは、比較的容易に入手できる。代表的なものとしては、UMLS Metathesaurus（生命・医学分野）、Wikipedia（カテゴリを意味クラスと見なすことができる）、Freebase（一般ドメイン）などが挙げられる．そこで、本研究では、比較的低コストで準備できる概念・実体の表現事例（語彙辞書）と、意味クラスの情報が付与されていない大量の生テキスト群（コーパス）を用い、意味クラスが付与された訓練データを自動獲得し、自動獲得された訓練データから固有表現抽出器を構築する。

[1] http://nlp.cs.nyu.edu/irex/index-j.html
[2] http://www.clips.ua.ac.be/conll2003/ner/
[3] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/
[4] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2008T04

ID	遺伝子名	別名	生物種
57126	CD177	NB1; PRV1	Homo sapiens

表 1: Entrez Gene のレコード例

(a) It is clear that in culture media of **AM**, cystatin C and cathepsin B are present as proteinase-antiproteinase complexes.

(b) Temperature in puerperium is higher in **AM**, lower in PM.

図 1: 辞書引きによる意味クラスのタグ付け例（タグ付けされた箇所を太字で表示）

2　提案手法

2.1　訓練データ構築

訓練データ構築を行うために、対象ドメインの文書、抽出したい意味クラスと、その表現を含む大規模語彙データベースを選定する。本研究では、生物医学分野の文献データベースである PubMed[5]の論文抄録（約1000万件）を対象ドメインとし、意味クラスとして遺伝子及びタンパク質名、語彙データベースとして Entrez Gene[6]を採用した。Entrez Gene は約680万件のレコードから構成され、各レコードには遺伝子ID、遺伝子名、タンパク質名、正式名称、生物種、詳細説明等が記載されている（表1）。今回の実験では、生物種が人間（*Homo sapiens*）である Entrez Gene レコードに限定し、遺伝子名、タンパク質名、正式名称、別名から表現事例を抽出し、辞書を作成した。

[5] http://www.ncbi.nlm.nih.gov/pubmed
[6] http://www.ncbi.nlm.nih.gov/gene

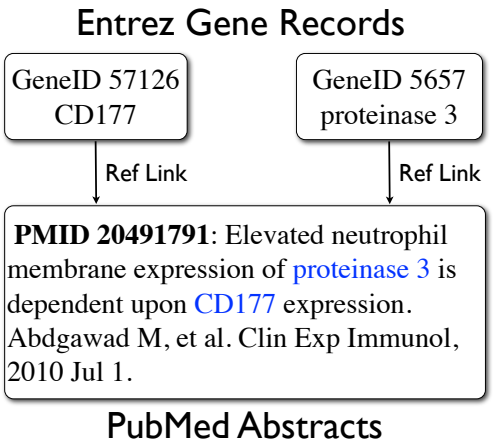

図 2: 関連 PubMed 文献へのリンク例

このように獲得した辞書に対し、PubMed の論文抄録を最長一致探索などで照合すれば、PubMed の論文抄録中に対して、意味クラスのタグを自動的に付与できる。しかし、このように辞書引きでタグ付けを行うと、遺伝子やタンパク質ではない表現を間違ってタグ付けしてしまうことがある。たとえば、図 1 は、Entrez Gene に含まれる表現「AM」がコーパス中で出現する箇所に、自動タグ付けを行った結果の例である。図 1 (a) における「AM」は、タンパク質の一種であるのでタグ付け結果が正しいが、(b) における「AM」は「午前」の意味で用いられており、太字の箇所をタンパク質名としてタグ付けすることは不適切である。これは、コーパス中の単語の意味の曖昧性を考慮していないためで、略称が多く用いられる生命・医学系の文献では、一般的な語に間違って意味クラスを付与してしまうことがある。

本研究では、表現の曖昧性問題を回避するために、Entrez Gene の各レコードが提供している参考文献情報を用いることにした。図 2 に、Entrez Gene が収録している参考文献情報の例を示した。この例では、#57126（表現例は「CD177」）と#5657（表現例は「proteinase 3」）の 2 つの Entrez Gene レコードを説明するための参考文献として、#20491791 の PubMed 論文抄録が挙げられている。論文抄録中で各表現は 1 つだけの意味を持つと仮定すれば、#20491791 の論文抄録における「CD177」や「proteinase 3」という表現は、それぞれ#57126 と#5657 の Entrez Gene レコードに言及していると考えられるため、これらの表現は遺伝子・タンパク質名である可能性が高い。本研究では、各 Entrez Gene レコードから参照されている論文抄録に対してのみ、対応する表現の自動タグ付けを行い、自動タグ付けの適合率を向上させた。

- ... in the following order: *tna*, **gltC**, **gltS**, *pyrE*; *gltR* is located near ...
- The three genes concerned (designated *entA*, *entB* and **entC**)
- Within the hypoglossal nucleus large amounts of **acetylcholinesterase** (*AChE*) activity are ...

図 3: Entrez Gene により参考文献として挙げられていなかったため、タグ付けから漏れてしまった例

$S \leftarrow (w_1, ..., w_{|S|})$ トークンのベクトル
$I \leftarrow \{i | 1 \leq i \leq |S|, w_i$は前節の手法により意味クラスが付与された $\}$
$C \leftarrow \{, \ . \ ; \ : \ (\ )\ and\ or\}$ 記号と接続詞の集合
$A \leftarrow \emptyset$ 正解クラスタグの集合
while $I \neq \emptyset$ **do**
  $i \leftarrow I$ から要素を *pop* する
  $A \leftarrow i$ を *push* する
  if $i \leq |S| - 2 \wedge w_{i+1} \in C \wedge w_{i+2}$は意味クラスタグが付いていない $\wedge\ w_{i+2} \in D$ **then**
    $I \leftarrow \{i + 2\}$ を *push* する
  end if
  if $i \geq 3 \wedge w_{i-1} \in C \wedge w_{i-2}$は意味クラスタグが付いていない $\wedge\ w_{i-1} \in D$ **then**
    $I \leftarrow \{i - 2\}$ を *push* する
  end if
end while

図 4: 等位構造タグ付けアルゴリズム

2.2 訓練データ拡張

前節の手法によりタグ付けを行った場合、本来遺伝子・タンパク質名であるはずであるが、Entrez Gene から参照されていないため、タグ付けできない事例が増える。図 3 に、前節の手法によりタグ付けされた表現（太字で表示）と、本来タグ付けされるべき表現（斜体で表示）を示した。前節の手法によりタグ付けされた表現の周辺に着目すると、「tna」「pyrE」「gltR」「entA」「AChE」といった語もタグ付けされるべきと推察されるが、これらの表現を収録している Entrez Gene レコードが、図 3 の文献を参照していないため、タグ付けされなかった。これは、Entrez Gene レコードの参考文献リンクが、レコードを説明するための参考情報という位置づけであり、網羅性が保証されていないためである。

本研究では、対象とする意味クラスに属さない表現

名前	詳細	実際の値の例
w	トークン	Human
wl	小文字化	human
pos	品詞	NNP
chk	チャンクタグ	B-NP
shape	文字種パターン	ULLLL
shaped	文字種パターン 2	UL
type	文字タイプ	InitCap
$p_n(n = 1...4)$	接頭辞 n 文字	(H,Hu,Hum,Huma)
$s_n(n = 1...4)$	接尾辞 n 文字	(n,an,man,uman)

表 2: 機械学習に用いた特徴

にタグ付けしてしまう可能性を抑えつつ、タグ付けする表現の網羅性を改善するため、等位構造を図 4 のアルゴリズムで解析し、タグ付けの表現を拡充した。このアルゴリズムは、前節の手法でタグ付けされた表現から、「,」「.」「and」などの等位接続を示しうるトークンを経由して到達できる表現が、Entrez Gene に（参考文献を考慮せずに）含まれているのであれば、タグ付けを拡張するというものである。このルールを適応することで、前節で構築したタグ付けデータの適合率を落とさずに、再現率を改善することが期待される。

2.3 機械学習

自動的に獲得した訓練データに対して、機械学習を用いて固有表現抽出器を構築した。訓練データに対して、GENIA tagger[7]を適用し、トークン切り出し、品詞タグ付け、チャンキングを行った。前節までの手法を用い、固有表現が出現している箇所には、IOB2 記法を用いてラベル付けを行った。機械学習アルゴリズムとしてサポートベクトルマシンを用い、文の先頭から末尾に向けて、トークンのラベルを一つずつ順に推定した。すなわち、文のトークン列 $x_1, ..., x_T$ に対して、以下の予測を $t = 1$ から T まで繰り返すことで、トークンのラベル列 $y_1, ..., y_T$ を求めた。

$$y_t = \underset{y}{\operatorname{argmax}}\, s(y|x_t, y_{t\ \ 1}), t \in \{1, ..., T\}$$

ただし、$s(y|x_t, y_{t\ \ 1})$ はサポートベクトルマシンが x_t のラベルを y と予測するときのスコア（素性の重みの和）である。サポートベクトルマシンの実装としては、liblinear[8]を用い、one-vs-the-rest 法により多クラス分類問題に拡張した。

表 2 に、学習に用いた特徴を挙げた。あるトークン（表 2 の例では「Human」）に対して、トークン文字列（w）、小文字化したトークン文字列（wl）、品詞（pos）、チャンクタグ（chk）、トークンの文字種パターン（shape）、文字種パターンから同一の文字種を間引きしたもの（shaped）、文字種タイプ（type）、トークンの接頭辞（p_n）、トークンの接尾辞（s_n）の特徴を取り出している。このうち、トークンの文字種パターン（shape）とは、トークン中に含まれる文字を大文字（U）、小文字（L）、数字（D）などの記号に縮退させたもの、shaped は shape のパターン文字列の中で同一の記号が連続する部分を一つにまとめたものである。文字タイプ（type）とは、「先頭が大文字で始まる」「全部が大文字で書かれている」「全部が数字である」「記号を含む」などの条件式にマッチする場合に発火する特徴である。本研究では、現在位置のトークンに対して、前後 2 トークン中に含まれる特徴量のユニグラム、及びバイグラム（但し wl, p_n, s_n は除く）を用いて素性を構成した。また、直前のトークンのラベルを現在位置のトークンの素性として用い、CRF で用いられるラベルバイグラム素性（遷移素性とも呼ばれる）を擬似的に導入した。

[7] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/

[8] http://www.csie.ntu.edu.tw/ cjlin/liblinear/

手法	P	R	F1
辞書マッチング	39.03	42.69	40.78
ref なし+svm	22.81	24.63	23.68
ref+svm	67.01	38.57	48.96
ref+拡張+svm	57.31	53.32	55.24

表 3: 各手法で作成した訓練データによる評価実験結果

3 実験と結果

提案手法で構築した固有表現抽出器を、BioNLP 2009 Shared Task[9]の Genes and Gene Products (GGP) コーパスで評価したときの適合率（P）、再現率（R）、F1 スコア（F1）を、表 3 に載せた。この評価では、固有表現抽出器が予測した固有表現の境界と、GGP コーパスの固有表現の境界が完全に一致する場合のみ、正解と見なしている。今回構築した固有表現抽出器が対象とする意味クラスと、GGP コーパスが対象とする意味クラスはオーバーラップが多いと考えられるが、GGP コーパスはアノテーションの指針をもって構築されているため、提案手法のタグ付けの基準と一致しない恐れがある。したがって、評価スコアが低く出る傾向にあるが、提案手法の性能の目安を調べたり、手法の各要素の貢献を調べるには、GGP コーパスで十分であると考えた。

[9] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/SharedTask/

表 3 では、GGP コーパスに対して辞書マッチングのみを行うベースライン手法（辞書マッチング）、参考文献情報を用いずに単なる辞書引きのみで学習データを構築したもの（ref なし+svm）、参考文献情報を用いて学習データを構築したもの（ref+svm）、等位構造に対してタグ付けの拡張を行って学習データを構築したもの（ref+拡張+svm）の実験結果を示している。各手法を PubMed 全体に適用したときに得られる訓練データ量は異なるが、今回の実験では比較のため、トークン数が同数（約 455 万トークン）となるよう調整している。

GGP コーパスに辞書マッチングを適用するベースライン手法の F スコアは、40.78 であった。機械学習を用いた固有表現抽出手法は、この辞書引きのみのベースライン手法の性能を上回る必要がある。しかしながら、Entrez Gene の参考文献を用いずに学習データを構築した場合、F スコアが 23.68 となり、ベースライン手法を下回ってしまった。これは、2.1 節でも説明したように、曖昧性の高い表現に対して間違ったタグが付与された学習データを構築してしまったためだと考えられる。

これに対し、Entrez Gene の参考文献情報を利用して学習データを構築した場合、F スコアは 48.96 へと改善され、ベースライン手法を上回ることができた。参考文献情報を用いると、適合率が大幅に改善されており（22.81 → 67.01）、学習データのタグ付け誤りが削減されることが分かる。さらに、2.2 節の手法により、自動タグ付けの結果を等位構造を持つ表現に拡張すると、適合率がやや低下するものの、再現率が大幅に向上し（38.57 → 53.32）、F スコアは 55.24 となった。適合率がやや低下していることから、タグ付け誤りが増加したと考えられるが、F スコアが向上していることから、2.2 節の手法の有効性を示すことができた。

4 関連研究

これまで、ラベル無しテキストから訓練データを自動獲得し、意味クラスタガーを学習させる研究 [6] や、辞書やデータベースを用いてラベル無しテキストをタグ付けし学習に用いる研究 [1, 7] が行われている。生物医学分野の文献における遺伝子の同定問題へのアプローチとして、生物種データベースを用いて訓練データを構築し、分類器を学習するアプローチ [4] もある。大規模情報を用いた固有表現抽出では Web 上のデータから精度の高い訓練データを生成する試み [5] もなされている。我々の手法は、これらの研究より大規模な語彙データベースとラベル無しテキストを用いること、語彙データベースが収録している付加情報として参考文献リンク情報を積極的に利用している点に特徴がある。

5 結論

本論文では、語彙データベースとラベル無しコーパスから自動構築した学習データを用い、固有表現抽出器を構築する手法を述べた。生命・医学文献を対象とし、Entrez Gene の参考文献情報を用いることにより、学習データの精度を改善できることが分かった。提案手法による学習データの自動構築では、タグ付け漏れを完全に防ぐことができないため、今後は半教師有り学習の適用 [2, 3] などを進めていきたい。

参考文献

[1] Kedar Bellare and Andrew McCallum. Learning extractors from unlabeled text using relevant databases. In *Sixth International Workshop on Information Integration on the Web*, 2007.

[2] Ruihong Huang and Ellen Rilo . Inducing domain-speci c semantic class taggers from (almost) nothing. In *Proceedings of ACL2010*, pp. 275–285, 2010.

[3] Zornitsa Kozareva. Bootstrapping named entity recognition with automatically generated gazetteer lists. In *Proceedings of EACL2006: Student Research Workshop*, pp. 15–21, 2006.

[4] Alexander A. Morgan, Lynette Hirschman, Marc Colosimo, Alexander S. Yeh, and Je B. Colombe. Gene name identi cation and normalization using a model organism database. *Journal of Biomedical Informatics*, Vol. 37, No. 6, pp. 396 – 410, 2004.

[5] Casey Whitelaw, Alex Kehlenbeck, Nemanja Petrovic, and Lyle Ungar. Web-scale named entity recognition. In *Proceeding of CIKM2008*, pp. 123–132, 2008.

[6] 村本英明, 鍛冶信, 末永直樹, 喜連川優. ラベルなしデータからの意味カテゴリタガーの学習. 第 5 回 NLP 若手の会シンポジウム, 2010.

[7] 土田正明, 水口弘紀, 久寿居大, 大和田勇人. 辞書とタグ無しコーパスを用いた固有表現抽出器の学習法. 第 23 回人工知能学会全国大会, 2009.