Vol.2011-NL-201 No.11
Vol.2011-SLP-86 No.11
2011/5/17

固有表現抽出のための大規模訓練データの自動獲得

宇佐美　佑[†1]　Han-Cheol Cho[†1]
岡崎直観[†2]　辻井潤一[†3]

固有表現抽出は，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おり，人名，地名，組織名，遺伝子名など，様々な意味クラスで試みられている．高い性能をもつ固有表現抽出器を構築するためには，あらかじめ意味クラスを付与した訓練データを用意し，機械学習アルゴリズムに基づいて構築するのが一般的である．しかしながら，訓練データの整備は，人手での作業に頼っているのが現状である．これでは，様々なドメイン・意味クラスで，広く固有表現抽出を利用しようにも，訓練データの入手性が固有表現抽出器構築のボトルネックになると考えられる．そこで，本研究では，より入手の容易な語彙データベースと生テキストを用いることで，固有表現抽出のための訓練データを人手に依らず自動的に獲得する手法を提案する．語彙データベースに含まれる豊富な情報を利用することで，高適合率な訓練データを自動獲得し，等位構造解析と self-training を適用することで，人手で作成した訓練データに迫る，高品質な訓練データを獲得した．

Automatic Acquisition of Huge Training Data for Named Entity Recognition

Yu Usami,[†1] Cho Han-Cheol,[†1] Naoaki Okazaki[†2]
and Jun'ichi Tsujii[†3]

†1 東京大学大学院情報理工学系研究科
Graduate School of Information Science and Technology, The University of Tokyo
†2 東北大学大学院情報科学研究科
Graduate School of Information Sciences, Tohoku University
†3 マイクロソフトリサーチ アジア
Microsoft Research Asia

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る．固有表現抽出は，単純には以下のような辞書マッチング問題として実装することができる．

(1) 抽出したい意味クラスの実体・概念の表現のリスト（gazetteer）を用意する．
(2) システムに与えられた文書を先頭から走査し，文書中の表現が辞書に含まれていた場合は，固有表現とみなす．

このアプローチは一見すると上手くいくように思えるが，いくつかの根本的な問題を抱えている．第一に，実体・概念の表現リストは，対象とする意味クラスに属するすべての表現を網羅する必要がある．しかしながら，固有表現抽出が対象とする意味クラスは，人物名，組織名，場所，遺伝子名，タンパク質名，病名など多岐にわたり，それぞれの分野において日々新語が生まれているため，現実的な設定とは言えない．第二に，辞書に含まれる表現に多義性がある場合，文書中で用いられている表現の意味が，辞書の意味カテゴリと一致するか，調べなければならない．つまり，文書中の表現そのものでは対象意味クラスの表現かどうかを決定することができず，前後の文脈を考慮して意味クラスの付与する必要がある．

このような辞書マッチングによる手法が抱える問題を克服するために，近年ではサポートベクトルマシン（SVM）や条件付き確率場（CRF）など，機械学習アルゴリズムを用いて固有表現抽出器を構築する研究が盛んである[1)]．機械学習に基づく固有表現抽出では，与えられた表現が，対象意味クラスに含まれるか否かを判定する分類問題として定式化される．このような分類モデルは教師あり学習で獲得するのが一般的であり，そのためには人手で意味クラスを付与した訓練データが必要になる．訓練データの整備は，対象ドメインの文書から対象意味クラスの表現を適切に見分けることができる専門家が，手作業で行わなければいけないため，費用や時間の面で多大なコストがかかる．したがって，固有表現抽出のための訓練データは特定の意味クラス・ドメインでしか整備されておらず，そのデータ量も限られているのが現状である．

本稿では，大規模な語彙データベースと，大量の生テキストコーパスを利用することで，固有表現抽出のための訓練データを自動的に獲得する手法を提案する．本研究で得られた知見は，以下の通りである．

(1) 単純な辞書マッチングで訓練データを自動獲得しても，訓練データの質が悪く，高性能な固有表現抽出器を構築できない．

Vol.2011-NL-201 No.11
Vol.2011-SLP-86 No.11
2011/5/17

(2) 語彙データベースに含まれる参考文献情報を使うことで，訓練データを高適合率・低再現率にすることができ，固有表現抽出器の性能も大幅に改善された．

(3) (2) で得られた訓練データの再現率を改善するため，等位構造解析と self-training を適用したところ，いずれも固有表現抽出器の性能が改善された．

(4) 最終的に自動獲得した訓練データは，人手で作成された訓練データに，固有表現の F1 スコアで 5.23 及ばなかったが，人手作業に迫る高品質な訓練データを獲得できることが示せた．

2. 提案手法

提案手法は，語彙データベースと生テキストコーパスを用いて，訓練データを自動的に獲得する手法である．本研究では，語彙データベースとして Entrez Gene[⋆1]を用いる．Entrez Gene は遺伝子とタンパク質のデータベースであり，約 680 万件のレコードそれぞれに，正式名称，別名，生物種名，詳細説明等が記載されている．図 **1** は Entrez Gene のレコード例である．提案手法では，これらのレコードより，正式名称，別名を集め，Entrez Gene 内に記載されている全ての遺伝子とタンパク質からなる辞書を構築した． また，生テキストコーパスとしては，2009 年版 MEDLINE[⋆2]の全体を利用することとした．MEDLINE は生物医学分野の論文抄録のデータベースであり，約 1,000 万件の論文抄録テキストが収録されている．これらの言語資源を用い，本研究では生物医学分野の文書から遺伝子名，タンパク質名を抽出する．

訓練データから固有表現抽出器を構築する手順は，以下の通りである．はじめに，GENIA tagger[⋆3]を適用することで，訓練データを，スペース，ハイフン (-)，コンマ (,)，ピリオド (.)，セミコロン (;)，コロン (:) で区切られた文字列（トークンと呼ぶ）に分割し，品詞（POS）タグやチャンクタグを付与する．固有表現のセグメントをラベルで表現するため，IOBES 記法[2] を採用した．

固有表現抽出は，与えられた文書のトークンごとに意味クラスの IOBES ラベルを与える多値分類問題と定式化することができる．今回は，線形カーネルの SVM の二値分類を，one-vs-the-rest 法により多値分類に拡張したものを学習アルゴリズムとして用いた．本研究で用いた SVM は，文の t 番目のトークン x_t が与えられたとき，以下のようにラベル y_t を予測する．

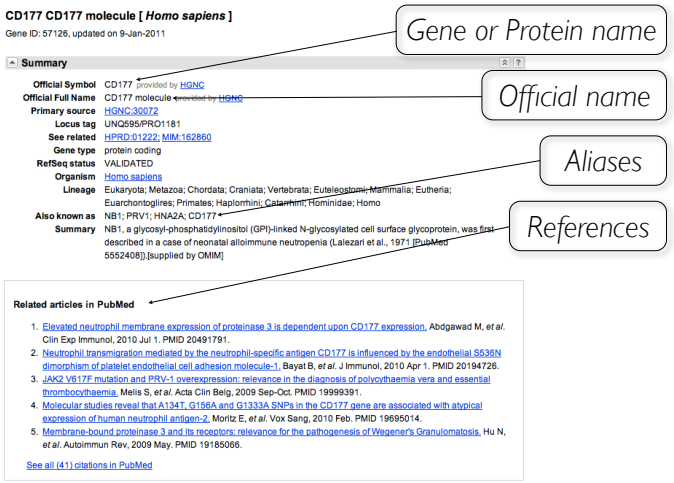

図 **1** Entrez Gene のレコード例
Fig. 1 Entrez Gene record sample.

$$y_t = \underset{y}{\operatorname{argmax}}\, \mathrm{score}(y|x_t, y_{t-1})$$

上式において，$\mathrm{score}(y|x_t, y_{t-1})$ はトークン x_t がラベル y であるスコア（特徴量の重みの和）を表す．y_t の予測において y_{t-1}（前のトークンの予測ラベル）を用いることで，CRF におけるラベル・バイグラム素性を擬似的に導入した．文が x_1 から x_T までのトークンからなるとき，文の先頭（y_1）から文の末尾（y_T）にかけて，上式のラベル予測を順に行う．本研究では，SVM の実装として liblinear[⋆4]を用いた．

表 1 は SVM での学習時に用いた素性のリストである．それぞれのトークン（表 1 では “Human” を例にした）に対し，次のような素性を作成した：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

⋆1 http://www.ncbi.nlm.nih.gov/gene
⋆2 http://www.ncbi.nlm.nih.gov/MEDLINE
⋆3 http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/
⋆4 http://www.csie.ntu.edu.tw/~cjlin/liblinear/

Vol.2011-NL-201 No.11
Vol.2011-SLP-86 No.11
2011/5/17

表 1　機械学習において用いた素性例
Table 1　Example of features used in machine learning process.

素性名	説明	例
w	トークン文字列	Human
wl	小文字化したトークン文字列	human
pos	品詞	NNP
chk	チャンクタグ	B-NP
shape	トークンの文字種パターン	ULLLL
shaped	トークンの文字種パターン 2	UL
type	文字種タイプ	InitCap
$p_n (n = 1...4)$	トークンの接頭辞	(H,Hu,Hum,Huma)
$s_n (n = 1...4)$	トークンの接尾辞	(n,an,man,uman)

タイプ（type），トークンの接頭辞（p_n），トークンの接尾辞（s_n）．トークンの文字種パターン（shape）とは，トークン中の文字を大文字（U），小文字（L），数字（D）に縮退したものである．文字種パターンから同一の文字種を間引きしたもの（shaped）は文字種パターン（shape）に似ているが，連続する同一文字種を一文字に縮めている．例えば，表 1 のように “ULLLL”（shape）ならば “UL”（shaped）となる．文字種タイプ（type）とは，そのトークンが「大文字で始まる」,「すべて大文字である」,「すべて数字である」,「記号を含む」などの，特定の条件を満たすかどうかを表す．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素性のユニグラム，及びバイグラム（但し wl，p_n，s_n は除く）を用いて特徴を構成した．

2.1　予備実験—単純な辞書マッチングによる訓練データの自動獲得

予備実験として，2009 年版 MEDLINE 全体に対して，Entrez Gene より構築した辞書の単純なマッチングを行うことにより，約 9 億トークンの訓練データを自動獲得した．獲得した訓練データを全て用いて学習することは，空間計算量の関係で不可能であったので，うち 440 万トークンの訓練データを学習に用いて，固有表現抽出器を構築した．この固有表現抽出器の性能を，BioNLP 2011 Shared Task[⋆1] EPI コーパスを用いて評価した．EPI コーパスは，訓練用データと開発用データが現在公開されているが，評価にはこれらを合わせた全体を用いた．EPI コーパスにおいて，付与されている意味クラスは GGP（Gene or Gene Product，遺伝子または遺伝子生成物）である．

表 **2** に，自動獲得した訓練データを用いて学習し，構築した固有表現抽出器の性能を示した．表 2 から分かるように，この方法で獲得した訓練データでは，学習をしても高性能な固有表現抽出器を構築できない（F1 スコア 14.27）．それどころか，Entrez Gene より構築した辞書を用いて，評価文書に直接辞書マッチングを適用した方が，良い結果となった（F1 スコア 40.78）．なぜこのように学習が効果的でないのかを明らかにするため，自動獲得した訓練データにどのように意味クラスが付与されているか調べた．

⋆1 https://sites.google.com/site/bionlpst/

表 **2**　予備実験結果
Table 2　Results of preliminary experiment.

手法	A	P	R	F1
評価文書に対する辞書マッチングによる抽出	92.09	39.03	42.69	40.78
自動獲得訓練データで学習した固有表現抽出器	85.76	10.18	23.83	14.27

(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 **2**　辞書マッチングによる意味クラス付与例（斜体の表現に意味クラスが付与されている）
Fig. 2　Dictionary-based gene name tagging example (tagged words are shown in italic typeface).

図 **2** は獲得された訓練データの一部である．例（a）における *AM* という語は，遺伝子名であるので正しい意味クラス付与である．しかし，例（b）における *AM* は，遺伝子名でもタンパク質名でもなく，午前を表す *ante meridiem* の略語なので偽陽性（false positive）である．このように，特に略語や頭字語に対して，単純な辞書マッチングによる意味クラス付与は，偽陽性を与える可能性が多いという問題を抱える． 自動獲得した訓練データには，同様な誤った意味クラス付与が非常に多く発見された．訓練データの質が悪いと，上記の曖昧性の問題を解消できない固有表現抽出器が構築され，十分な性能を発揮することができない．学習データの質の問題を，機械学習の側から解決するのは非常に困難であるので，本研究では学習データの質を改善する方法を考える．

2.2　参考文献情報を用いた訓練データの自動獲得

自動獲得する訓練データの適合率を改善するため，Entrez Gene の各レコードが収録している参考文献情報を利用した．図 **3** は参考文献情報の例であり，Entrez Gene に含まれる *AM* という遺伝子のレコードの参考文献情報を示している．参考文献情報とは，その

Vol.2011-NL-201 No.11
Vol.2011-SLP-86 No.11
2011/5/17

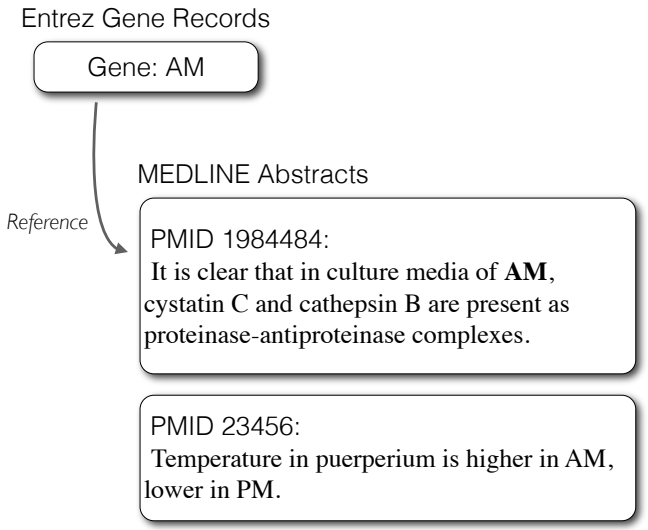

図 3　MEDLINE 論文抄録への参考文献情報例
Fig. 3　Reference to MEDLINE abstract example.

レコードの記述する遺伝子やタンパク質が記述されている文献として，MEDLINE の文献 ID（PMID）を示すものである．2.1 節における誤った意味クラス付与は，Entrez Gene に *AM* という遺伝子名のレコードが存在するため，全ての *AM* という表現を遺伝子と認識してしまったことで引き起こされたのである．参考文献情報を見ると，この *AM* のレコードは MEDLINE #1984484 の文書しか参照していない．そこで，Entrez Gene の各レコードから参照されている MEDLINE の論文抄録においてのみ，辞書マッチングによる自動アノテーションを行うことにし，意味クラスの表現を誤ってアノテートするケースを軽減することにした．こうすることで，MEDLINE #23456 の文書で *AM* へ遺伝子の意味クラスを付与しなくなり，図 2 の例（b）の偽陽性を解消できる．

参考文献情報を用いた，訓練データの自動獲得は次のように行う．

(1)　Entrez Gene より正式名称，別名，参考文献情報を集め，表現のリストと参考文献情報が紐付けられた辞書を構築する．
(2)　MEDLINE 文書全体に対し，辞書マッチングを適用する．
(3)　マッチした表現の参考文献情報が，該当する MEDLINE の論文抄録を参照している場合のみ，意味クラスを付与する．

このプロセスを MEDLINE 全体に適用したところ，4,800 万トークンの訓練データが自動獲得でき，うち 300 万トークンに意味クラスが付与された．

2.3　訓練データ拡張

2.2 節では，参考文献情報を用いて高適合率の訓練データを獲得する方法を述べた．しかしながら，この手法を用いたことで獲得できる訓練データは，偽陰性の多い（低再現率な）ものとなってしまう．図 **4** で斜体で記述された部分は，全て遺伝子名である．この中で，下線が引かれた表現に関しては，Entrez Gene に対する辞書引きより見つかったレコードにおいて，該当する MEDLINE の論文抄録が参照されていたため，2.2 節の手法で遺伝子名の意味クラスが与えられた．しかし，下線が引かれていない表現に関しては，Entrez Gene に対する辞書引きで見つかるものの，そのレコードにおいて，該当する MEDLINE の論文抄録が参照されていないため，2.2 節の手法では遺伝子名の意味クラスを付与できなかった．このように，斜体下線無しの表現は，意味クラスが付与されるべきだが付与できていないもの（偽陰性）となってしまい，学習の妨げとなってしまう．このようなことが起こるのは，Entrez Gene のレコードに含まれる参考文献情報に網羅性が保証されていないためである．

適合率を維持したまま再現率を改善するため，本研究では等位構造に着目した．すなわち，2.2 節の手法でアノテートされた名詞と等位の関係にある名詞は，同じ意味クラスに属すると考えた．図 5 に，等位構造解析に基づく意味クラス付与の拡張アルゴリズムを示した．このアルゴリズムでは，2.2 節の手法で意味クラスが付与された表現から等位構造を表す記号（“,”，“.”，“and” 等）を経て到達できる表現が，Entrez Gene から（参考文献の制約を無視して）辞書引きで見つかった場合に，意味クラスを付与する．

2.4　self-training

2.3 節の手法は，等位構造に基づいて訓練データの再現率を改善する手法であるため，偽陰性の問題がすべて解決出来るわけではない．そこで，獲得した訓練データに存在する偽陰性を自動的に修正するために，本研究では self-training を導入することにした．一般的に self-training とは，少量の正しく意味クラスの付与されたデータ（シードと呼ばれる）と，意味クラス付与のされていない大量の生テキストコーパスを用い，次に挙げる操作を繰り返しながら訓練データを獲得する[3]．

(1)　シードより分類モデルを構築し，そのモデルを生テキストコーパスに適用する．
(2)　適用した結果，固有表現であると新しく認識された表現に意味クラスを付与する．

Vol.2011-NL-201 No.11
Vol.2011-SLP-86 No.11
2011/5/17

- ... 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 4　偽陰性の例
Fig. 4　False negative examples.

Input: Sequence of sentence tokens S, Set of symbols and conjunctions C, Dictionary without reference D, Set of annotated tokens A
Output: Set of Annotated tokens A

begin
for $i = 1$ to $|S|$ **do**
  if $S[i] \in A$ **then**
    $j \leftarrow i - 2$
    while $1 \leq j \leq |S| \wedge S[j] \in D \wedge S[j] \notin A \wedge S[j+1] \in C$ **do**
      $A \leftarrow A \cap \{S[j]\}$
      $j \leftarrow j - 2$
    end while
    $j \leftarrow i + 2$
    while $1 \leq j \leq |S| \wedge S[j] \in D \wedge S[j] \notin A \wedge S[j-1] \in C$ **do**
      $A \leftarrow A \cap \{S[j]\}$
      $j \leftarrow j + 2$
    end while
  end if
end for
Output A
end

図 5　等位構造解析に基づく意味クラス付与の拡張アルゴリズム
Fig. 5　Coordination analysis algorithm.

Input: Labeled training data D, Machine learning algorithm A, Iteration times n, Threshold θ
Output: Trained model M_n

begin
Current labeled training data $T_0 \leftarrow$ initial size data from D
Construct base model M_0 with T_0 by means of A
$i \leftarrow 0$
$D \leftarrow D \backslash T_0$
while $i \neq n$ **do**
  $U \leftarrow$ some amount of size data from D
  $L \leftarrow$ Annotated U with model M_i
  $S \leftarrow$ Selected new labeled data over θ form L
  $U_{new} \leftarrow$ Apply S labels to U
  $T_{i+1} \leftarrow T_i \cup U_{new}$
  $M_{i+1} \leftarrow$ Construct new model with T_{i+1}
  $D \leftarrow D \backslash U$, $i \leftarrow i + 1$
end while
Output M_n
end

図 6　self-training アルゴリズム
Fig. 6　Self-training algorithm.

(3)　新しく意味クラスの付与された文を，シードに加える．

これらのプロセスを，定められた反復回数分繰り返すか，生テキストを使い切るまで続け，大量の訓練データを獲得する．

本研究では，大量の訓練データを既に獲得しているため，self-training の問題設定とは異なる．2.3 節の手法で獲得した訓練データの適合率は高いと考えられるので，すでにアノテーションされた箇所は信頼し，まだアノテーションされていない箇所に固有表現があるかどうか，検討したい．**図 6** は，self-training アルゴリズムに改良を施し，偽陰性の可能性がある表現への意味クラス付与を行うものである．まず，2.3 節で獲得した訓練データ（D）を，シードデータ（T_0）と残りのデータ（$D \setminus T_0$）に分ける．その後，$0 \leq i \leq n$ について以下の操作を繰り返す．

(1)　シードデータ（T_i）より分類モデル（M_i）を構築する．
(2)　残りのデータの一部を取り出す（U）．
(3)　モデル（M_i）を取り出したデータ（U）に適用する．
(4)　適用した結果，固有表現であると認識された表現に意味クラスを付与する．
(5)　新しく付与された表現のうち，確信度が閾値（θ）を超えるものと，2.3 節の手法でアノテートされた表現を統合したデータ（U_{new}）を，シードデータ（T_i）に加える．

本研究では，初期シードデータサイズを 68 万トークン，各反復で残りの訓練データより取り出すデータサイズを 22 万トークンとした．

新しく意味クラスを付与する際に，無差別に追加していては訓練データの質が低下するおそれがあるため，アノテーションの認定に確信度（Confidence）[4] を用いた．トークン x のラベルが SVM により y と予測され，そのスコア（特徴量の重みの和）が $\mathrm{score}(x, y)$ と計算されるとき，トークン x に対する予測の確信度（$\mathrm{Confidence}(x)$）は，次の式で計算される．

$$\mathrm{Confidence}(x) = \mathrm{score}(x, y) - \max(\forall_{z \neq y} \mathrm{score}(x, z))$$

つまり確信度とは，予測した最上位のラベルのスコアと，次に高かった予測ラベルのスコアとの差（マージン）である．確信度は，個々のトークンへのラベル予測に対して計算されるものであるので，予測したラベルが単一のトークンの固有表現（IOBES 記法においては S）であった場合は，この確信度が閾値（θ）を超えれば，固有表現のアノテーションとして

Vol.2011-NL-201 No.11
Vol.2011-SLP-86 No.11
2011/5/17

認定する．もし，予測したラベルが複数トークンからなる固有表現（IOBES 記法において B または I または E）の場合は，個々のトークンへのラベル予測の確信度を平均した値が閾値（θ）を超えた場合に，その表現のアノテーションを認定する．すなわち，複数トークンからなる固有表現 x_i，...，x_j に対する確信度は，次のように計算される．

$$\text{Confidence}(x_i, ..., x_j) = \frac{1}{j-i+1} \sum_{k=i}^{j} \text{Confidence}(x_k)$$

図 6 の self-training アルゴリズムによる学習の際にも，2 節冒頭で説明した素性を採用する．しかし，本研究における self-training に期待することは，固有表現であるのに意味クラスが付与されていないと思われる表現に対して，意味クラスを付与することである．したがって，意味クラスが付与されないというルールを，それぞれのトークンの表現から学習してしまうことを避けたい．そこで，本研究の self-training アルゴリズムによる学習の際には，現在位置のトークンに対し，ユニグラムのトークン文字列（w）の特徴（現在位置のトークン自身）を素性から削除することにし，各トークンの文脈（対象表現の周辺の語句）を重視して学習するようにした．

3. 実験と結果

本節では，提案手法により獲得した訓練データを用いて，固有表現抽出器を構築し，性能を評価する．提案手法を MEDLINE 全体に適用したところ，4,800 万トークンから成る訓練データが得られた．今回の実験では，計算機資源の制約から，全体の 10%のデータを訓練データとして用いることにした．評価文書としては，2.1 節と同じく，BioNLP 2011 Shared Task EPI コーパスを用いた．評価尺度としては，精度（A），適合率（P），再現率（R），F1 スコア（F1）の 4 つの尺度を用いた．それぞれの固有表現の予測が正しいかどうかは，固有表現のセグメント境界が左右共に厳密に一致した時に限り，正解とする．

3.1 提案手法評価

2 節において，訓練データを自動獲得する際に用いる 3 つの手法を提案した．それぞれの手法の適用段階における訓練データで，固有表現抽出器を構築し，性能を測定したものを表 3 に載せた．

表 3 の先頭行「辞書マッチング」は，評価文書に対して単純な辞書マッチングを行った結果であり，2.1 節の評価結果を再掲したものである．ここでの F1 スコア 40.78 が，評価における基準値となる．表 3 の二番目以降は，すべて獲得した訓練データを用いて機械学習をし，固有表現抽出器を構築したうえで，評価文書への適用した結果を示している．「訓練データ自動獲得 (参考文献情報なし)」は，2.1 節で行った，単純な辞書マッチングによって自動獲得した訓練データを用いた固有表現抽出器の結果である．ここから，参考文献情報を利用することで，大きく固有表現抽出器の性能が向上した（「＋参考文献情報」）．適合率は最高値（69.25%）を達成し，再現率は低い（39.12%）ものの，基準値を上回る F1 スコア 50.00 となった．さらに，等位構造解析をして訓練データを拡張することで（「＋等位構造解析」），適合率の低下は抑えつつ（-2.46%），再現率が大きく改善し（+8.32%），F1 スコアは 55.47 に向上した（+5.47）．最後に，self-training を行うことで（「＋ self-training」），適合率は若干下がるものの（-3.07%），再現率は改善し（+3.74%），F1 スコアは最も高い 56.77 となった（+1.30）．

表 3　評価結果
Table 3　Results of evaluation.

	手法	A	P	R	F1
学習なし	辞書マッチング	92.09	39.03	42.69	40.78
学習あり	訓練データ自動獲得（参考文献情報なし）	85.76	10.18	23.83	14.27
	＋参考文献情報	93.74	**69.25**	39.12	50.00
	＋等位構造解析	93.97	66.79	47.44	55.47
	＋ self-training	**93.98**	63.72	**51.18**	**56.77**

self-training の各反復による性能向上を詳しく見るため，反復毎に固有表現抽出器を構築し，F1 スコアをプロットしたものが図 **7** である．反復する毎に F1 スコアが向上していくことが分かり，最後の 22 回目の反復まで性能の向上が続いた．なお，表 3 の実験結果は，他の実験と訓練データサイズを揃えるために，反復を 17 回まで繰り返した訓練データを使用している．

3.2 人手で作成された訓練データとの比較

現在の最高水準の固有表現抽出器は，人手で作成された訓練データを用いたものがほとんどである．ここでは，本研究で得られた自動獲得した訓練データで学習した固有表現抽出器と，人手で作成された訓練データで学習した固有表現抽出器を比較する．また性能の比較とともに，自動獲得した訓練データは，人手で作成された訓練データのどの程度の分量に匹敵するのかを調べる．評価文書として，BioNLP 2011 Shared Task EPI コーパスを用いた．今回，教師あり学習モデルの固有表現抽出器を構築するにあたって，EPI コーパスの訓練用データで学習し開発用データのみを評価に用いることとした．同様に，提案手法の固有表現抽出器も，開発用データのみで評価する．

Vol.2011-NL-201 No.11
Vol.2011-SLP-86 No.11
2011/5/17

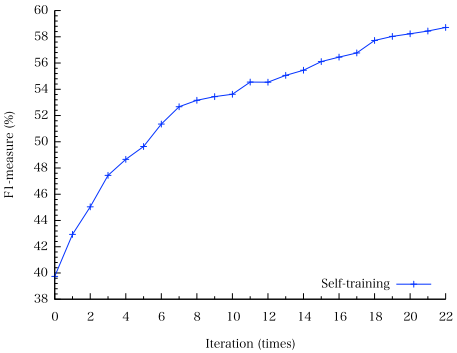

図 7　self-training 結果
Fig. 7　Results of self-training.

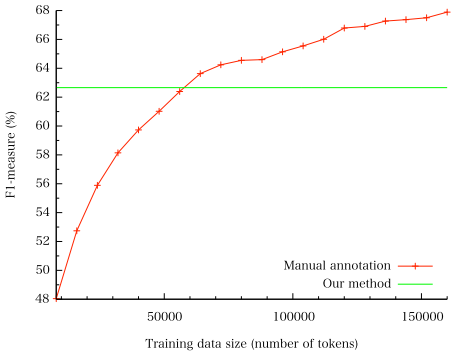

図 8　人手で作成した訓練データとの比較
Fig. 8　Manual annotation vs. our method.

EPI コーパスの訓練用データを，全体の 20 分の 1 サイズ毎に増やしていき，それぞれの訓練データ量において学習した固有表現抽出器の性能を，図 **8** にプロットした．提案手法の固有表現抽出器の性能として，self-training の反復を 22 回行った訓練データを用いて，構築した固有表現抽出器の F1 スコア 62.66 を横線で示している．人手で作成された訓練データで学習した固有表現抽出器の性能は，全体を用いると F1 スコア 67.89 となった．本研究において自動獲得した訓練データは，人手で作成された訓練データに F1 スコアで 5.23 及ばなかったが，人手で作成された訓練データの約 40%（6 万トークン，2,000 文）に匹敵することが確認できた．

3.3 考　　察

提案手法によって自動獲得した訓練データは，人手で作成された訓練データに及ばなかったものの，高い性能をもつ固有表現抽出器を構築できることが分かった．ここで，さらに高い性能を目指すために，提案手法に不足している点を考察する．

まず，等位構造解析のみでは解消しきれない，偽陰性の例が確認できている．

tna loci, in the following order: *tna*, *gltC*, *gltS*, *pyrE*; *gltR* is located near ...

上記の例において，斜体で記述された部分は，全て遺伝子名を表す．等位構造解析によって，下線の引かれた表現より周囲の斜体の表現まで意味クラス付与をすることができた．しかし，冒頭の斜体太字の表現は，現状の等位構造解析では意味クラスを付与することができない．このような場合にも，正しく意味クラスを付与するために，one sense per discourse 法[5] などの導入を検討する必要がある．

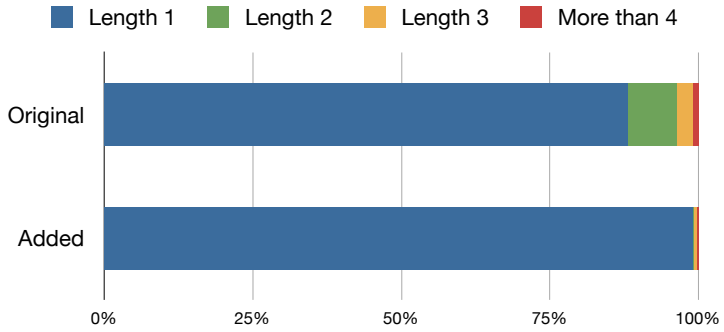

図 9　トークン長分布
Fig. 9　Distribution of entity length.

self-training を用いても，F1 スコアの改善が 1.0 未満に留まった．self-training がそれほど効果的でなかった原因は様々だと考えられるが，self-training において追加された固有表現のトークン長の分布（図 **9**）を調べることで，追加される固有表現のトークン長に偏りが生じていることが分かった．図 9 には，固有表現のトークン長の分布が，self-training を用いた場合の分布（Original）と，用いなかった場合の分布（Added）で記されている．この図からわかるように，提案手法の self-training において，単一トークンからなる固有表現ばかりが追加されており，複数トークンからなる固有表現が追加されていない．こうして，元の分布に比べ，偏った訓練データになっており，単一トークンからなる固有表現ばかりを予測する固有表現抽出器になっている可能性が考えられる． 追加する固有表現のトークン長分布を変質させる，本研究の self-training における固有表現の追加の仕方（Confidence の平均と閾値の比較）に問題があると考えられる．

4. 関 連 研 究

本研究は，高性能な固有表現抽出器を，人手作業なしで構築しようということを目的としている．このように，人手に頼らない方針の研究として，少量のシードから訓練データを獲得しようとしたり，シードすら用いない方針を採用している研究がいくつか存在する．Vlachos と Gasperin は，少量のシードデータから bootstrapping[4] を用いて訓練データを獲得し，生命医学分野テキストにおける固有表現抽出器の構築を行った[6]．対象ドメインは

Vol.2011-NL-201 No.11
Vol.2011-SLP-86 No.11
2011/5/17

違うが，シードデータから固有表現抽出のための大規模な訓練データを獲得した研究としては，Whitelaw らがシードデータと Web データからの，大規模訓練データの獲得に成功した[7]．また，Kozareva は，固有表現抽出器を二つの分類器で別々に学習させ，性能を向上させる手法を用いた[8]．シードデータも用いずに，訓練データを獲得し固有表現抽出器を構築する試みは，村本らが Wikipedia と blog を用いて取り組んだ[9]．これは，人手の作業を最小限にしつつ訓練データの獲得を行っているが，固有表現抽出器の構築や評価までは行えていない．

5. 結　　論

本稿では，固有表現抽出のための訓練データを自動的に獲得する手法を提案し，評価実験を行った．実験の結果，提案手法は単純な辞書マッチングによる固有表現抽出よりも高い性能を発揮した．人手で作成された訓練データには及ばなかったが，参考文献情報を用いた高適合率の達成，等位構造解析と self-training による低再現率の改善と，全ての提案手法が高性能な固有表現抽出器の構築に有効であることが示せた．

しかし，将来的な課題はいくつか考えられる．提案手法の self-training アルゴリズムでは固有表現の追加の際に，トークン長の分布が偏ってしまった．この固有表現の追加方法を，より適切なものに改良する必要がある．大規模な訓練データを獲得することができたが，空間計算量の制約から訓練データ全てを利用することが出来なかった．オンライン学習の可能な SVM を実装することで，獲得した訓練データ全てを学習に利用できるようにしたい．本研究では，Entrez Gene の参考文献情報が十分でないために，等位構造解析や self-training を導入した．そもそも，参考文献情報が網羅的であれば，これらの手法を用いなくても，高適合率かつ高再現率な訓練データが獲得できる可能性がある．不足している参考文献情報を予測する問題として実装し，より良い訓練データを獲得するアプローチを試みたい．本研究においては，提案手法を生物医学分野での遺伝子名とタンパク質名の抽出に用いたが，提案手法は特定のドメイン・意味クラスに依存した手法ではないと考える．異なるドメイン・意味クラスでも，提案手法が有効であることを示したい．

参 考 文 献

1) Nadeau, D. and Sekine, S.: A survey of named entity recognition and classification, *Lingvisticae Investigationes*, Vol.30, No.1, pp.3–26 (2007).
2) Ratinov, L. and Roth, D.: Design challenges and misconceptions in named entity recognition, *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, CoNLL '09, pp.147–155 (2009).
3) Zadeh Kaljahi, R.S.: Adapting self-training for semantic role labeling, pp.91–96 (2010).
4) Huang, R. and Riloff, E.: Inducing domain-specific semantic class taggers from (almost) nothing, *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, Association for Computational Linguistics, pp.275–285 (2010).
5) Gale, W.A., Church, K.W. and Yarowsky, D.: One sense per discourse, *Proceedings of the workshop on Speech and Natural Language*, HLT '91, Association for Computational Linguistics, pp.233–237 (1992).
6) Vlachos, A. and Gasperin, C.: Bootstrapping and evaluating named entity recognition in the biomedical domain, *Proceedings of the HLT-NAACL BioNLP Workshop on Linking Natural Language and Biology*, LNLBioNLP '06, Association for Computational Linguistics, pp.138–145 (2006).
7) Whitelaw, C., Kehlenbeck, A., Petrovic, N. and Ungar, L.: Web-scale named entity recognition, *Proceeding of the 17th ACM conference on Information and knowledge management*, CIKM '08, ACM, pp.123–132 (2008).
8) Kozareva, Z.: Bootstrapping named entity recognition with automatically generated gazetteer lists, *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, EACL '06, Association for Computational Linguistics, pp.15–21 (2006).
9) 村本英明，鍛冶伸裕，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).