点予測と能動学習を用いた固有表現抽出の提案

小林　滉河†　　若林　　啓††

† 筑波大学情報学群知識情報・図書館学類　〒 305–8550 茨城県つくば市春日 1-2
†† 筑波大学 図書館情報メディア系　〒 305–8550 茨城県つくば市春日 1-2
E-mail: †s1511506@u.tsukuba.ac.jp, ††kwakaba@slis.tsukuba.ac.jp

あらまし　近年では様々な専門分野に対して，固有表現認識を行う研究が増えている．しかし，専門分野におけるアノテーションコーパスは一般分野のコーパスに比べて非常に少ない．その上，専門ドメインのアノテーションには専門的知識が要求される為，一般ドメインに比べてコーパスの作成に大きなコストがかかる．そのため専門ドメインにおいて，少ないアノテーションコーパスで高い性能での固有表現抽出を実現させる手法が求められている．本研究は一部の箇所に対してタグを付与出来る部分的アノテーション コーパスを利用可能な点予測を用いることで，アノテーションが少ない状況においても高い性能を実現できる固有表現抽出手法と能動学習への適用を提案する. 提案手法の有用性を示すため，提案手法と既存手法である条件付き確率場と深層学習モデルに能動学習を適応させ，同一のテストデータに対しての性能を計測した．

キーワード　自然言語処理，固有表現認識，点予測，能動学習

1 はじめに

固有表現抽出（Named Entity Recognition; NER）とは，文章の中から人名・地名・組織名といった固有名詞や時間表現，金額表現等の語句を抽出する自然言語処理の技術のことであり，自然言語処理における情報抽出タスクの一つとして知られている．抽出された固有表現は情報検索や質問応答といった様々なタスクに応用され，精度を大きく左右させる要因になる．多くの固有表現抽出では訓練データとして，アノテーションコーパスを利用し機械学習モデルを学習する手法が用いられる．アノテーションコーパスとはテキストに対して人間が言語的な情報を付与したデータのことである．またテキストに情報を付与する人のことはアノテータと呼ばれる．固有表現抽出におけるアノテーションコーパスの表現方法として，IOB2 や IOE2，IOBES といった様々なタグのフォーマットが存在するが，本研究では IOB2 フォーマットを用いる．IOB2 フォーマットは二つの要素の組を各単語に付与することで固有表現を表す．一つ目の要素では（Begin; B）を固有表現の先頭，（Inside; I）を固有表現の続き，（Other; O）を固有表現以外として表すことで固有表現かそうではないかを示す．二つ目の要素ではその固有表現が何に分類されるかを示す．組織名なら ORG（ORGANIZATION），人名なら PER（PRESON）といった分類先を表す単語が後に続く．O にはこのような要素は付与されない．このような分類はコーパスやタスクごとに定義される．図 1 は「田中」を人名（PER），「筑波大学」を組織名（ORG）として IOB2 フォーマットに基づきラベル付けを行なった場合である．

図 1 のように全ての単語列 $\boldsymbol{x}$ に対して，ラベル列 $\boldsymbol{y}$ が付与されているコーパスのことはフルアノテーションコーパスと呼ばれる．対して，図 2 のように一部の単語にしかラベルが付与されていないコーパスのことを部分的アノテーションコーパスという．

単語列 $\boldsymbol{x}$	田中	は	筑波	大学	の	学生	だ	.
ラベル列 $\boldsymbol{y}$	B-PER	O	B-ORG	I-ORG	O	O	O	O

図 1　フルアノテーションコーパスの例

単語列 $\boldsymbol{x}$	田中	は	筑波	大学	の	学生	だ	.
ラベル列 $\boldsymbol{y}$		O	B-ORG	I-ORG	O			O

図 2　部分的アノテーションコーパスの例

近年では食材名や調理方法などの固有表現を持つ料理ドメインや，物質材料ドメイン，バイオドメインといった様々な専門分野に対して固有表現抽出を行う研究が増えている．しかし，専門分野におけるアノテーションコーパスは一般分野のコーパスに比べて非常に少ない．その上，専門ドメインのアノテーションには専門的知識が要求される為，一般ドメインに比べてコーパスの作成に大きなコストがかかる．そのため専門ドメインにおいて，少ないアノテーションコーパスで高い性能での固有表現抽出を実現させる手法が求められている．

少数のアノテーションコーパスで効果的な学習を行うアプローチとして能動学習が知られている．能動学習はアノテーションコーパスを作成する際に，大量のアノテーションなしデータから機械学習モデルによって，より効果的に学習が進むインスタンスに対して，アノテータにラベルを付けてもらう手法である．既存の固有表現抽出手法の多くは訓練データとしてフルアノテーションコーパスしか利用出来ない．そのため能動学習を適用させる際にアノテータに対して文章を推薦し，その中に含まれる全ての単語にラベルを付与してもらう必要がある．しかし，ラベルが付与されることで学習が効率的に進む単語というのは文章中においてあまり多くはないと考えられる．例え

ば頻出する固有名詞への複数回に渡るアノテーションはモデルの学習に対してあまり効果的ではない．このようなモデルの学習に効果的ではない単語に対してのアノテーションはアノテータへの負担を増加させる原因の一つになっている．

本研究では，点予測 [1] を用いた固有表現抽出手法と能動学習への適用を提案する．点予測は現在主流の手法である深層学習を用いた固有表現抽出に比べ，フルアノテーションコーパスが大量にある状況において性能は劣る．しかし，点予測では部分的アノテーションコーパスが利用可能であり，文章中の全単語に対してアノテーションを行う必要がない．そこで提案手法では部分的アノテーションコーパスを訓練データとして利用可能である点予測のメリットを最大化するために能動学習を適用する．実験では既存手法に対して能動学習を適用させたモデルと提案手法を比較し，少ないアノテーションコストで，高い抽出性能を持つことを示す．

本稿は次のとおりに構成される．第 2 章では，固有表現抽出と能動学習の関連研究を紹介する．その際，固有表現抽出タスクに対して能動学習を適用した論文についてもいくつか紹介する．第 3 章では点予測による固有表現抽出と能動学習への適用手法について説明する．第 4 章では点予測に対して能動学習が有効であるか示す実験を行なった後，既存手法と提案手法について比較実験を行う．第 5 章では，結論と本研究の今後の展開について議論を行う．

2 関 連 研 究

2.1 固有表現抽出

人名や組織名，地名といった固有表現を自動的に抽出する技術である固有表現抽出は，情報抽出技術の一つであり自然言語処理の基礎技術である．一般的なドメインにおける固有表現コーパスとして新聞記事に対して地名や人名，組織名などの固有表現タグを用いてアノテーションを行なった CoNLL-2003 [2] というコーパスがよく知られている．Florian ら [3] は CoNLL-2003 の固有表現抽出コンペティションにおいて複数個の機械学習モデルの組み合わせを用いることで，F1 値 88.76%という高い性能を実現した．McCallum ら [4] は固有表現抽出を系列ラベリングタスクとみなし，条件付き確率場（Conditional Random Fields; CRF）を利用して固有表現抽出を行った．近年では Ronan ら [5] が単語列に対して畳み込みニューラルネットワーク（Convolutional Neural Network; CNN）を用いた手法を提案した．以降，深層学習を用いた固有表現抽出が主流となっている．Huang [6] らは Ronan らのモデルの CNN エンコーダを Bidirectional Long Short-Term Memory（双方向 LSTM）エンコーダに置き換えたモデルを提案した．Lample ら [7] は文字や単語レベルの情報を双方向 LSTM を利用してモデル化し，文字レベルの情報を CNN を利用することで，今まで人手で設計していたデータの前処理を必要とせずによりよい性能を達成した．Ma ら [8] は双方向 LSTM と CNN，CRF を組み合わせたモデルを提案し，CoNLL-2003 において 91.21%の F1 値の精度を達成した．しかし，これらの手法はフルアノテーションコーパスを訓練データとした固有表現抽出モデルの構築を目標としており，部分的アノテーションコーパスを利用することが出来ない．そのためアノテータは単語列全てに対してアノテーションをしなければならず，モデルがあまり必要としていない単語に対してもラベルを付けるコストが発生する．また周辺尤度を用いることで部分的アノテーションコーパスが利用可能な条件付き確率場 [9] も提案されているが学習にかかる時間が非常に長く，何度も学習を行う必要がある能動学習を適用させるのは現実的ではない．

2.2 能 動 学 習

少量のアノテーションコーパスでよりよい精度を出す方法として能動学習が知られている．

能動学習では機械学習モデルを用いて，ラベルなしデータの中から情報が多そうなインスタンスを選択する．そして選択したインスタンスをラベルの答えを知っているオラクル（典型的にはアノテータ）に対してラベルを問い合わせる．問い合わせによりラベルが付与されたインスタンスを訓練データに加え，更にモデルの学習を行う．これらの一連の処理を繰り返すことで，能動学習は有益なデータの収集を優先的に行うことを目的としている．能動学習のシナリオにはいくつかの種類が存在する [10] が，本研究ではプールベース能動学習（Pool-based Sampling）について検討する．プールベース能動学習は，大量のラベルなしデータが取得済みの状況において，モデルの学習に有益なデータを選択する．また，プールベース能動学習のようなシナリオでは各インスタンスの情報量尺度の評価を元に問い合わせを行う．情報量尺度の計算を行うためのクエリ戦略は数多く提案されている．

Uncertainty Sampling [11] はモデルからみて最も予測が不確かなラベルをアノテータに問い合わせる手法であり，その不確かさとして扱われる指標には様々な手法が存在する．その中でも本稿の実験にて利用した指標について紹介する．

まず，最もシンプルな指標である Least Confident は $\boldsymbol{x}$ を単語列，$\boldsymbol{y}^*$ を事後確率が最も高いラベル列としたとき，以下のように定義される．

$$\phi^{LC}(\boldsymbol{x}) = 1 - P_\theta(\boldsymbol{y}^*|\boldsymbol{x}) \tag{1}$$

つまり Least Confident では現在のモデルにおいて，事後確率が最も小さいと予測したラベル列 $\boldsymbol{y}^*$ を持つ単語列 $\boldsymbol{x}$ を推薦する指標である．また $|\boldsymbol{x}| = |\boldsymbol{y}^*| = 1$ のとき，多クラス分類モデルに対する指標としてみなすことができる．この指標を点予測では用いた．

次に二つの最も確信度が高いと予測したラベルの差からクエリ戦略を行う方法である Marginal Sampling という指標について説明する．$\boldsymbol{y}_1^*$ を事後確率が最も高いラベル列，$\boldsymbol{y}_2^*$ を事後確率が二番目に高いラベル列としたとき，以下のように定義される．

$$\phi^{M}(\boldsymbol{x}) = -\big(P_\theta(\boldsymbol{y}_1^*|\boldsymbol{x}) - P_\theta(\boldsymbol{y}_2^*|\boldsymbol{x})\big) \tag{2}$$

Marginal Sampling は Least Confident とは異なり，最も確からしいラベル列以外の曖昧性についても考慮出来るようになるため得られる情報量が増えると考えられている．

別のクエリ戦略として Query-By-Committee（QBC）[12] が存在する．このクエリ戦略では複数個のモデルからなる $\mathcal{C} = \{\theta^{(1)}, \ldots, \theta^{(C)}\}$ を用いる．commitee は各インスタンスに対してラベルを予測するが，その時インスタンスに対して異なるラベルが多く付いたものが最も有益なインスタンスだとみなす．QBC にも Vote Entropy [12] や Kullback Leibler [13] といった不一致度を評価するための指標が提案されている．またラベルが得られたときに，大きくモデルが変更されると考えられるインスタンスをアノテータに問い合わせる Expected Model Change [14] や，汎化誤差を小さくするようなインスタンスを問い合わせる Expected Error Reduction [15] といった数多くのクエリ戦略が研究されている．

2.3 固有表現抽出に対する能動学習の適用

能動学習を固有表現抽出タスクに適用させる手法は，これまでにいくつか提案されている．Settles ら [16] は系列ラベリングタスクを解くために，条件付き確率場に対して能動学習を適用させる手法の提案を行った．Yanyao ら [17] は認識すべき表現が多いとき，深層学習モデルにおいて出力層に CRF レイヤを用いた場合に比べて，LSTM レイヤを用いた方が高速であることを示し，文字 CNN，単語 CNN，LSTM を組み合わせた CNN-CNN-LSTM モデルを考案し，深層学習モデルであっても比較的高速に能動学習が行える手法を提案した．しかし，これらの手法は，学習にフルアノテーションコーパスを用いるモデルに基づいていることから，アノテータに対して文章に含まれる全ての単語についてのラベルを問い合わせる必要がある．そのためアノテータはモデルの学習にあまり効果的ではない単語に対してもラベルを付与する必要があり，コスト増加の原因となっている．

本研究では，高速に学習可能かつ部分的アノテーションコーパスを訓練データに利用できる点予測に対して能動学習を適用させることによって，アノテーションコストの削減を目指す．

3 手　　法

3.1 点予測による固有表現抽出

本研究では Neubig [18] が提案した点予測を固有表現抽出に活用出来るように拡張を行う．そのため本節では点予測について説明する．単語列を $\boldsymbol{x} = \langle x_1, x_2, \ldots, x_T \rangle$，ラベル列を $\boldsymbol{y} = \langle y_1, y_2, \ldots, y_T \rangle$ とする．点予測では部分的アノテーションコーパスから得られる情報を元にラベルの推定を行うために，条件付き確率場や深層学習モデルのようにタグの並びの尤もらしさを考慮しない．x_i に対して y_i は B-LOC, I-LOC, O のような IOB2 フォーマットのラベルを取ることから，y_i の推定を行うタスクは多クラス分類問題になる．多クラス分類問題を解くための手法としてロジスティック回帰（Logistic Regression）やサポートベクターマシン（Support Vector Machine; SVM），決定木（Decision Tree）などの機械学習モデルが存在するが，本研究では能動学習に適用するため，高速に学習と予測を行える多クラスロジスティック回帰を点予測で利用するモデルとして用いる．

多クラスロジスティック回帰は L 個のラベル $\mathcal{Y} \in \{1, 2, \ldots, L\}$ の中からラベル $y \in \mathcal{Y}$ を一つ推定するモデルである．多クラスロジスティック回帰では，以下の式を用いて入力 x に対してのクラスラベル y が属している確率を求める．

$$P(y|x) = \frac{\exp\left(\boldsymbol{w}_y^T \boldsymbol{f}(x)\right)}{\sum_{y' \in \mathcal{Y}} \exp\left(\boldsymbol{w}_{y'}^T \boldsymbol{f}(x)\right)} \tag{3}$$

$\boldsymbol{f}(x)$ は素性関数 $f(x)$ からなる素性ベクトルである．素性関数は人手によって設計される．また素性関数を記述する方法は素性テンプレートと呼ばれる．また $\boldsymbol{w}$ は素性ベクトルに対しての重みベクトルであり，学習の際にはこの重みベクトルを訓練データに適合するようにして調整を行う．

多クラスロジスティック回帰では，学習に用いる訓練データ $(\boldsymbol{x}, \boldsymbol{y}) = (\langle x_1, x_2, \ldots, x_n \rangle, \langle y_1, y_2, \ldots, y_n \rangle)$ を用いて対数尤度 l（log-likelihood）を最大化させることで重みベクトル $\boldsymbol{w}$ の最適化を行う．対数尤度 l は全ての訓練データの $P(y_{(i)}|x_{(i)})$ の積の対数として，以下の式で定義出来る．

$$l(\boldsymbol{w}) = \log \prod_{i=1}^{n} P(y^{(i)}|x^{(i)}; \boldsymbol{w}) = \sum_{i=1}^{n} \log P(y^{(i)}|x^{(i)}; \boldsymbol{w}) \tag{4}$$

この対数尤度は凸関数であるため，勾配降下法を用いることで効率的に重みベクトル $\boldsymbol{w}$ を計算することが出来る．

点予測による固有表現抽出ではタグ予測の対象である単語を x_i と m を窓幅としたとき，x_i の周辺単語である $x_{i-m+1}, \ldots, x_{i-1}$ と $x_{i+1}, \ldots, x_{i+m}$ を入力として，これらの素性を元にラベル y_i を予測する（図 3）．

つまり多クラスロジスティック回帰を提案手法におけるタスクに書き直すと以下のような式になる．

$$P(y|x_{i-m+1}, \ldots, x_{i+m}) = \frac{\exp\left(\boldsymbol{w}_y^T \boldsymbol{f}(x_{i-m+1}, \ldots, x_{i+m})\right)}{\sum_{y' \in \mathcal{Y}} \exp\left(\boldsymbol{m}_{y'}^T \boldsymbol{f}(x_{i-m+1}, \ldots, x_{i+m})\right)} \tag{5}$$

Neubig [18] は日本語の単語分割に対して点予測を利用したが，本研究では英語の固有表現抽出に対して適用する．このため，単語 $x_{i-m+1}, \ldots, x_i, \ldots, x_{i+m}$ に対する素性として Neubig [18] は文字 n-gram ベースの素性を用いていたが，これを単語ベースの素性に変更する．以下が提案手法で定めた素性である．

- 単語の表層系: 識別するラベル y_i に対して窓幅 m を設定し，$x_{i-m+1}, \ldots, x_i, \ldots, x_{i+m}$ を素性として用いる．
- 単語種別: 単語に関する素性と同様に文字種に関する情報も $x_{i-m+1}, \ldots, x_i, \ldots, x_{i+m}$ から抽出し，素性とする．本実験では対象となる単語が大文字から始まっているか，大文字から構成されているかを単語種別から得られた素性として用いる．

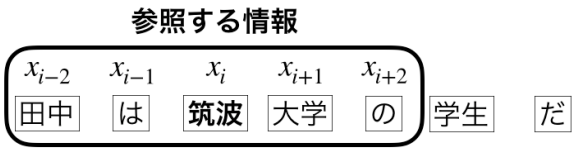

図 3　x_i のラベル推定の際参照する情報（$m = 2$ において）

- 品詞: ラベルの推定対象となる単語とその周辺単語の品詞も素性として利用する．

また森ら [1] は多値分類に対しての点予測に，各単語の品詞候補毎に分類器を作成し一対多方式（one-versus-rest）を用いることで品詞の推定を行なっている．しかし提案手法では全単語に対して一つの分類器を用いて固有表現抽出を行う．提案手法が単語毎に分類器を作成しなかった理由は未知単語に対するラベル付与の仕方である．森ら [1] は形態素解析のタスクに対して点予測を用いたため，学習コーパスや辞書に出現しなかった単語は名詞とみなすことが出来る．これに対して，固有表現抽出タスクでは未知の単語に対して O タグやその他特定のタグを付与すると，抽出性能が著しく減少することが予備実験によって分かった．そのため本稿では全単語に対して一つの多値分類器を使用する．

3.2 能動学習への適用

点予測では部分的アノテーションコーパスが利用可能であり，文章の一部の単語に対してのみラベルが付与されているデータも訓練データとして利用可能である．しかし，部分的アノテーションコーパスが利用可能でも似たような単語に対してばかりにラベルが付与されている場合学習はうまく進まない．モデルの学習にとって有益な単語に対して，優先的にタグ付けを行うアルゴリズムを提案する．

今回用いる能動学習にはプールベース能動学習を利用する．プールベース能動学習は少数のラベル付きデータ L と大量のラベルなしデータであるプール U があることを前提とした能動学習手法であり，現在の機械学習モデルの予測に基づいて，学習に最も有用なインスタンスをプールの中から選択し，アノテータに対してラベルの問い合わせを行う．アノテータはインスタンスに対してラベル付けを行い，それを元にモデルを更新する．これらの繰り返しによってモデルを学習を進行する．

また，一度に一つのデータについてのみラベル問い合わせを行ってモデルの再学習を行う能動学習は逐次型能動学習と呼ばれるが，この設定ではモデルが再学習を行なっている間アノテータの待機時間が生じてしまう．. このため，提案手法には一度に複数個のラベルなしデータを選択し，ラベルの問い合わせを行うバッチ型能動学習を利用する．

プールベース能動学習は一般的に Algorithm 1 で表される．このアルゴリズムではまずラベル付きデータ L に対し，関数 $train(\cdot)$ を用いてモデルを学習する．その後，第 2 章で説明したクエリ戦略 $\phi(\cdot)$ に基づき，アノテータに対して B 回プール U からラベル問い合わせを行う．問い合わせたデータはラベル付きデータとして L に加えられ，U から削除される．これを繰り返し行いモデルを学習していく．また Algorithm 1 は $B > 1$ のときバッチ型能動学習と，$B = 1$ のとき逐次型能動学習とみなすことができる．

Algorithm 1 点予測による能動学習

Require: ラベル付きデータ L, ラベルなしデータ U, クエリ戦略 $\phi(\cdot)$, バッチサイズ B

```
repeat
  θ ⇐ train(L)
  for b = 1 to B do
    x*_b ⇐ arg max_{x∈U} φ(x)
    L ⇐ L ∪ {⟨x*_b, label(x*_b)⟩}
    U ⇐ U − {x*_b}
  end for
until
```

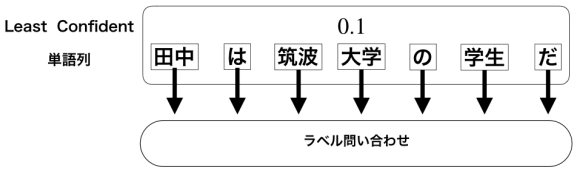

図 4　既存のモデルにおけるラベル問い合わせ

条件付き確率場や LSTM-CNN-CRF のような既存研究のモデルでは文章中の全体にタグがついていなければ訓練データとして扱うことが出来ない．そのため能動学習を適用した際，図 4 のように文章に含まれる単語全体に対してラベルの問い合わせを行う必要がある．しかし点予測に対して能動学習を適用すると図 5 のように，文章中の一部の単語だけを問い合わせることが出来る．このように同じ単語の数にアノテーションを行なっても，点予測では既存研究のモデルより重要だと予測する単語を多く問い合わせることが出来るため，効率的にモデルの学習が行えると考えられる．

アノテータへの問い合わせの方法について説明する．アノテータは，固有表現全体を確認しなければ，B タグや I タグの区別をしながらアノテーションを行うことはできない．このため，周辺の数単語を見せて，固有表現全体の範囲をアノテーションしてもらう方式を考える．これを単語単位の問い合わせとして形式的に述べると，問い合わせの手続きは以下のようになる．まず，クエリ戦略に基づいて単語 x_i が推薦される．次に，x_i が固有表現の途中である場合，固有表現全体にアノテーションが行えるように x_{i-1} のインスタンスをアノテータについて問い合わせる．この問い合わせを固有表現の始まりが出てくるまで続ける．また x_i が固有表現の始まり，または固有表現の途中であった場合，系列における一つ先のインスタンスである x_{i+1} について問い合わせを行う．これも固有表現の終わりが出現するまで行う．（図 6）

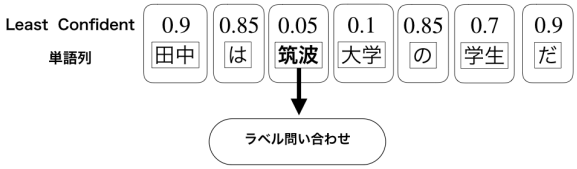

図 5　点予測におけるラベル問い合わせ

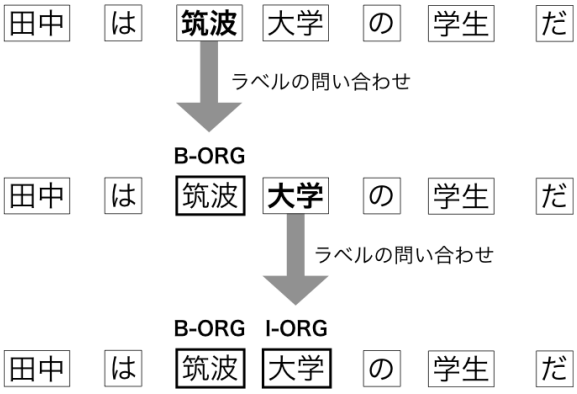

図 6　ラベル問い合わせの過程の例

4　実験と評価

4.1　学習データ

本研究では固有表現データセットとしてよく知られている CoNLL-2003 English [2] を利用した．CoNLL-2003 には training ファイルが一つと test ファイルが二つ用意されている．本研究では CoNLL-2003 コンペティションのレギュレーションに従い，各ファイルを training/validation/test の三つにデータを分割した．モデルの学習を行うための training データは少量のラベルありデータ L と大量のラベルなしデータであるプール U の二つに分割し，少量のラベルありデータ L だけを最初の学習に利用した．深層学習モデルのハイパーパラメータの探索には validation データセットを利用し，評価の際には test データに対する固有表現抽出の性能を計測した．

4.2　実 験 方 法

本研究では，点予測を用いた固有表現抽出における能動学習の有効性について既存手法との優位性を示すために二種類の実験を行う．一つ目の実験では点予測を用いた固有表現抽出に対して，能動学習が実際に有効であるかの検証を行なった．大量のラベルなしデータ U の中からランダムに単語を選択しラベルを問い合わせることで学習を行う点予測モデルとクエリ戦略 Least Confident [19] を用いて，ラベルなしデータの中から最も有益だと判断をした単語に対してラベルの問い合わせを行なった点予測モデルの性能を F1 値によって比較を行なった．

二つ目の実験では，既存手法に素直に能動学習を適用した場合との性能の比較を行なった．既存手法である条件付き確率場と，Ma ら [8] が提案した LSTM-CNN-CRF による固有表現抽出モデルに対し，それぞれ能動学習を適用させた．これらの手法と，提案手法である点予測による固有表現抽出モデルの性能を，F1 値によって比較を行った．LSTM-CNN-CRF のハイパーパラメータとして学習時のミニバッチサイズは 10，エポック数は 50 と設定した．また単語の分散表現には Ma ら [8] の実験に従い，GloVe [20] の 50 次元を利用した．

単語列	田中	は	筑波	大学	の	学生	だ	.
正解ラベル	B-PER	O	B-ORG	I-ORG	O	O	O	O
予測ラベル	B-PER	O	O	O	O	O	O	O

図 7　評価方法の例

それぞれの実験において，能動学習のクエリ戦略には Lewis らの Uncertainty Sampling [11] を用いた．また指標として点予測，条件付き確率場には式 (1) で示される Least Confident [19]，系列全体の確率を算出するのが困難な LSTM-CNN-CRF モデルに対しては式 (2) で示される Marginal Sampling [21] を利用した．

いずれの手法においても，一度に約 5,000 単語の問い合わせを行うバッチ型能動学習の設定とした．問い合わせを行う際，点予測ではラベルなしデータ中の全単語からクエリ戦略に基づき，推薦を行った 5,000 単語に対してラベルを付与した．一方，条件付き確率場と LSTM-CNN-CRF モデルでは単語ごとにラベルを付けることが出来ない．そのため，文章ごとにラベルの付与を行い，ラベルが付与された単語が 5,000 個を超えた時点で，その際行われているバッチを終了させる．

今回の実験では人手によるアノテーションは利用せず，正解の分かっているコーパスを用いて仮想的な条件の元で能動学習を行なった，そのため常に問われた単語に対して正しいラベルが付与される．

評価の際には固有表現の始点と終点が完全に一致している場合を正解とみなし，固有表現ではない O タグの一致は正解として考慮していない．例えば図 7 のような単語，正解ラベルと予測ラベルがあったとき，「田中」についての固有表現抽出は（B-PER）で一致しているため正解とみなす．「筑波大学」は（B-ORG, I-ORG）のラベルに対して（O, O）と予測しているため，抽出に失敗している．そのため精度は 100%，再現率は 50%であり，F1 値は 66.7%と計算される．

4.3　実 験 結 果

4.3.1　点予測の固有表現抽出に対する能動学習の有効性

ランダムにラベルの問い合わせを行う点予測モデルと，能動学習を適用し Least Confident を用いてラベルの問い合わせを行う点予測モデルの比較を行なった．抽出性能と単語に対するアノテーション数の関係を図 8 に示す．能動学習を適用した点予測モデルではラベルなしデータの一割強に対してアノテーションを行うだけで，ランダムにラベルを問い合わせたモデルが全単語に対してラベルを付与したときと同等の抽出性能を実現した．したがって，点予測に対して能動学習を適用することは必要とするラベル付きデータの数を減らし，アノテータの負担を少なくすることに貢献できると思われる．

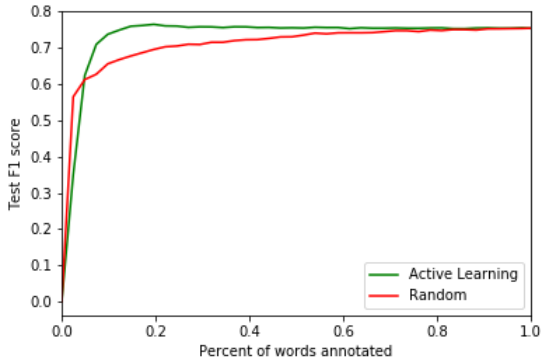

図 8　点予測のクエリ戦略の実験結果．x 軸はラベル付けを行なった単語の割合，y 軸は F1 値

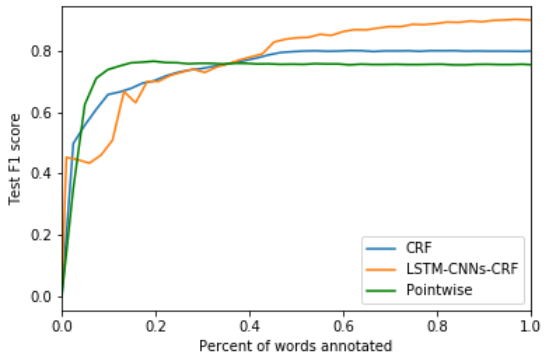

図 9　既存手法と提案手法の実験結果．x 軸はラベル付けを行なった単語の割合，y 軸は F1 値

4.3.2　能動学習を適用した固有表現抽出の性能比較

図 9 に既存手法と提案手法による固有表現抽出に対して能動学習を適用したときの抽出結果と単語に対するアノテーション数の関係を示した．提案手法である点予測は既存手法に比べ，ラベルありデータが少ない場合において高い抽出性能を得られた．また本実験では LSTM-CNN-CRF モデルに対してのみ Marginal Sampling を用いたが，提案手法である点予測を用いたモデルに対しても Marginal Sampling や更により多くの情報を得ることが出来る指標を適用することも可能であり，さらなる性能の向上も期待できる．しかし，提案手法ではアノテーションされた単語が約 40%を超えると他の手法に比べ抽出性能が劣っており，アノテーション数が増えたときの抽出性能の改善が今後の課題である．

5　結　　論

本研究では，点予測を用いた固有表現抽出と能動学習の適用手法の提案を行なった．部分的アノテーションが利用可能な点予測を用いることで既存の手法では不可能であった単語ごとのアノテーションが可能となった．また点予測に対して能動学習を適用させることによって，単語単位でアノテータにラベル問い合わせが出来るようになった．提案手法は深層学習を用いたモデルに比べ高速に学習を行えるため，バッチの大きさを小さくしても，既存手法と比較してあまりコストを増やすことなく，更に少ないアノテーションで高い性能にたどり付けることが期待できる．実験結果では無作為に選択した単語をアノテーションした点予測モデルに比べ，能動学習を適用した点予測モデルの方が少ないラベル付きデータで高い性能を得られることを示した．また既存手法である条件付き確率場や深層学習モデルに対して能動学習を適用した場合に比べても，提案手法の点予測モデルはラベル付きデータが少ない場合において，比較的高い抽出性能を達成できることを示した．

しかし，ラベル付きデータが増加するにつれ，既存手法に比べると性能が劣っていくという課題が残っている．能動学習を適用するかどうかに限らず，固有表現抽出タスクにおいて，十分なラベル付きデータが与えられる場合，点予測のようなモデルよりも，タグの並びの尤もらしさを考慮する条件付き確率場や深層学習モデルの方が抽出能力が高い．このため，部分的アノテーションコーパスを利用できる深層学習モデルについて今後考察していく必要がある．

文　　献

[1] 森信介, 中田陽介, Neubig Graham, 河原達也. 点予測による形態素解析. 自然言語処理, Vol. 18, No. 4, pp. 367–381, 2011.

[2] Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pp. 142–147, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[3] Radu Florian, Abe Ittycheriah, Hongyan Jing, and Tong Zhang. Named entity recognition through classifier combination. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pp. 168–171. Association for Computational Linguistics, 2003.

[4] Andrew McCallum and Wei Li. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pp. 188–191, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[5] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, Vol. 12, pp. 2493–2537, November 2011.

[6] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional LSTM-CRF models for sequence tagging. *CoRR*, Vol. abs/1508.01991, , 2015.

[7] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 260–270. Association for Computational Linguistics, 2016.

[8] Xuezhe Ma and Eduard Hovy. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1064–1074. Association for Computational Linguistics, 2016.

[9] Yijia Liu, Yue Zhang, Wanxiang Che, Ting Liu, and Fan

Wu. Domain adaptation for crf-based chinese word segmentation using free annotations. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 864–874, 2014.

[10] Burr Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

[11] David D Lewis and Jason Catlett. Heterogeneous uncertainty sampling for supervised learning. In *Machine Learning Proceedings 1994*, pp. 148–156. Elsevier, 1994.

[12] H. S. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, pp. 287–294, New York, NY, USA, 1992. ACM.

[13] Andrew McCallum and Kamal Nigam. Employing em and pool-based active learning for text classification. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, pp. 350–358, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.

[14] Burr Settles, Mark Craven, and Soumya Ray. Multiple-instance active learning. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pp. 1289–1296. Curran Associates, Inc., 2008.

[15] Nicholas Roy and Andrew McCallum. Toward optimal active learning through sampling estimation of error reduction. In *ICML*, 2001.

[16] Burr Settles and Mark Craven. An analysis of active learning strategies for sequence labeling tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pp. 1070–1079, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

[17] Yanyao Shen, Hyokun Yun, Zachary C. Lipton, Yakov Kronrod, and Animashree Anandkumar. Deep active learning for named entity recognition. *CoRR*, Vol. abs/1707.05928, , 2017.

[18] Graham Neubig. 点推定と能動学習を用いた自動単語分割器の分野適応. 言語処理学会年次大会, 2010, 2010.

[19] Aron Culotta and Andrew McCallum. Reducing labeling effort for structured prediction tasks. In *AAAI*, 2005.

[20] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *EMNLP*, Vol. 14, pp. 1532–1543, 2014.

[21] Tobias Scheffer, Christian Decomain, and Stefan Wrobel. Active hidden markov models for information extraction. In *International Symposium on Intelligent Data Analysis*, pp. 309–318. Springer, 2001.