Vol.2017-NL-234 No.4
2017/12/20

分野特有の教師なし固有表現認識

友利 涼[1,a)]　森 信介[2,b)]

概要：本稿では、教師なし手法による分野特有の NER を提案する。本手法は教師なし形態素解析手法である PYHSMM を拡張している。本手法は単語レベルの PYHSMM と semi-Markov CRF を統合しており、一般分野のコーパスによる擬似教師データ、対象分野の少数のシードデータを用いて NER を行う。本手法は、分野特有の固有表現コーパスが存在しない分野において、少ないコストで分野特有の教師なし NER を行うことができ、様々な分野や言語、固有表現体系に応用可能である。複数の分野において英語と日本語で実験を行い、提案手法は NER の精度を向上させることが確認できた。

キーワード：固有表現認識、教師なし学習、分野特有

Domain-Specific Unsupervised Named Entity Recognition

Suzushi Tomori[1,a)]　Shinsuke Mori[2,b)]

Abstract: In this paper, we propoce an unsupervised named entity recognition for a particular domain. Our method is based on PYHSMM which has been proposed for unsupervised word segmentation and pos tagging model. Our model incorporates word level PYHSMM and semi-Markov CRF and requires pseudo labeled data in general domain and seed data in a target domain. Our method can perform domain-specific NER with low cost in a domain which doesn't have domain-specific NE corpus. Our method can be applied to various domains regardress of languages and NE definitions. In the experiments, we took several target domains in English and Japanese as examples. Experimental results showed that our method improves the NER accuracy.

Keywords: Named Entity Recognition, Unsupervised Learning, Domain-Specific

1. はじめに

一般に固有表現認識 (NER) とは、テキスト中から人名や地名、組織名などを抽出する技術のことで、情報検索 [1] や関係抽出 [2][3]、共参照解析 [4] などに応用される。これらのタスクは構造化されていないデータをコンピュータで扱いやすくすることを目的とし、より高度な言語処理技術に用いられる。また、近年ではバイオ医療分野のテキストにおける DNA 名や protein 名などの認識を目的とするバイオテキスト NER [5] や料理レシピのテキストにおける食材名や道具名などの認識を目的としたレシピ NER [6] などの分野特有の固有表現体系が定義されたコーパスが提案され、それぞれ文献分類や手順書理解 [7] などの基礎技術となっている。一般の NER と同様に、それぞれの分野で、より高度な言語処理に必要となる。また、同じ分野でも応用面での違いから異なる固有表現体系が定義されることもあり、バイオ医療の分野では、疾患・治療関係認識 [8] のために疾患名や処置名などがラベル付けされたコーパス [9][10][11] などが存在する。

NER の研究の多くは機械学習の手法に基づいており、大量のラベル付きデータを用いることで高い精度を実現している [12][13]。しかし、大量のラベル付きデータが存在している分野は少なく、分野特有の NER では重大な問題

[1] 京都大学大学院情報学研究科
Graduate School of Informatics, Kyoto University
[2] 京都大学学術情報メディアセンター
Academic Center for Computing and Media Studies, Kyoto University
[a)] tomori.suzushi.72e@st.kyoto-u.ac.jp
[b)] forest@i.kyoto-u.ac.jp

になっている [14]。また、分野特有の固有表現コーパスの作成は、その分野の専門知識が必要となるためコストが大きく、異なる応用タスクや分野、言語ごとに固有表現コーパスを作成することは現実的ではない。そのため、多くの分野特有の NER はルールベース手法もしくはヒューリスティックな素性を大量に用いた機械学習モデルである [15]。これらの手法はどちらも特定分野に対する深い専門知識が必要となり、他分野への適用は難しい。

そこで、本研究では教師なし学習による分野特有の NER を提案する。本手法は Pitman-Yor Hidden Semi Markov Model (PYHSMM) による教師なし形態素解析 [16] を拡張した手法であり、分野特有のエンティティとそのクラスを同時に推定する。PYHSMM は分かち書きされていない文書から、言語モデルが最適化するような単語列とその品詞列をラベル付きデータなしで推定する。本手法では、対象となる分野のラベルなしデータと少数のシードデータ、一般分野のテキストを擬似教師データとして用いることで、対象となる分野にのみ多く出現する単語列を認識し、そのクラスを周辺文脈から推定する。

本手法は、ラベル付けがされていない分野のテキストに対し、少ないコストで分野特有の NER を行うことができ、従来の教師なし NER における特別な専門知識や用語辞書など用いずに、様々な分野や言語、固有表現体系に応用可能である。また、本手法は、特定の分野のテキストに対し、系列ラベリングを用いて事前に定義されたクラスの用語を取得する科学分野の情報検索 [17] や slot tagging [18] などにも応用可能である。

本稿の構成は以下の通りである。まず 2 節で教師なし NER の関連研究について述べ、3 節では教師なし形態素解析について簡単に説明する。4 節では提案手法について説明し、5 節で実験設定と実験結果について述べる。最後に 6 節で本稿をまとめる。

2. 関連研究

関連研究として、一般の NER のタスク説明とその代表的な解き方や分野特有の NER、教師なし NER について述べる。

本稿で行う NER は自然言語処理の重要な技術の一つである。一般的な NER は、対象を新聞記事、固有表現として人名クラス、地名クラス、組織名クラス等の 8 種類の固有表現クラスを扱った研究が広く行われている [19][20]。NER は様々な言語で行われており、英語やドイツ語 [19]、スペイン語やオランダ語 [21]、日本語 [22][23] などのコーパスが存在する。NER は、ある系列 (単語列) の各要素 (各単語) に適切なラベル列を付与する問題である系列ラベリング問題として解かれることが一般的である。隠れマルコフモデル [24] やサポートベクターマシン [25] や最大エントロピーモデル [26] などを用いた様々な手法が提案されており、条件付き確率場 (CRF) による系列ラベリング [27][28] がよく用いられる。また、近年では深層学習を用いて解く手法が提案されており、Bi-directional LSTM を用いて解く手法が CoNLL 2003 コーパス [19] において最高精度を記録している [13]。

分野特有の NER では、バイオ医療分野での研究が盛んである。GENIA コーパス [5] は、大量のバイオ医療分野の論文を整理するために提案され、DNA 名クラスや protein 名クラスなど 5 種類の分野特有の固有表現クラスを定義されている。CRF とヒューリスティックな素性を用いて解く手法 [15] やヒューリスティックな素性や分散表現などを用いてエンティティ検出をする手法 [29] などがある。また、文献整理以外にも様々な応用があり、バイオ医療分野の NER は、その分野における意味関係抽出 [30] などの基礎技術となっている。バイオ医療分野では GENIA コーパス以外でも、言語や応用ごとに様々なコーパスが作成されている。例えば、文献 [9][10][11] では、疾患名クラスなどの固有表現クラスが定義されており、疾患・治療関係認識 [8] の基礎技術となっている。他にも製品名やブランド名の認識 [31]、科学文書におけるタスク名やプロセス名の認識 [17] なども分野特有の NER といえる。その他にも様々な分野での固有表現コーパスが定義されているが、そのほとんどのコーパスにおいてラベル付きデータの少なさが問題になっている。文献 [14] は、転移学習と約 6,000 文程度のラベル付きデータを用いて分野特有の NER の精度向上を示した。

教師なし NER の研究では、ルールベースの手法による NER やエンティティが与えられた状態で少数のシードデータを元にクラス推定を行う研究が多い。文献 [32] は、ルールベースの手法を用いて会社名の認識を行った。文献 [33] は用語辞書とルールベースの手法を用いて、200 クラスの NER を提案した。文献 [34] と文献 [35] は少数のシードデータとヒューリスティックを用いてブートスラップ手法により、与えられたエンティティを事前に定義されたクラス (人名クラス、地名クラス、組織名クラスなど) に分類した。しかし、これらの手法はクラス推定のみを行っており、人手でヒューリスティックなルールを作成しているため、多数のクラス分類は行うことは難しく、特定の分野でのルール作成は専門知識が問われる。文献 [36] はバイオ医療分野において、注釈付きテキストを用いずに NER を行う手法を提案した。この手法はエンティティの認識とそのクラス推定を別々に行っている。まず、名詞句を取り出し、シードデータと TF-IDF スコアからエンティティを取得し、医学用語辞書を利用し分類を行う。この手法は医学用語辞書のように体系化された用語辞書の存在を仮定しており、すべての分野に適応できるものではない。

3. 教師なし形態素解析

本節では教師なし形態素解析について説明するため、まずNested Pitman-Yor Language Model (NPYLM) による教師なし単語分割 [37] について述べ、その後、NPYLM と CRF (NYCRF) による半教師あり単語分割 [38] と PYHSMM について述べる。

NPYLM[37] は分かち書きされていない文を単語 n-gram 言語モデルが最適 (予測力最大) となる単語列へ分割する手法である。これは階層的 Pitman-Yor 言語モデル (HPYLM)[39] を拡張することで文字列から言語モデルを構築し、言語モデルを最適化するような単語列を得ることができる。

HPYLM は Pitman-Yor 過程に基づくノンパラメトリックベイジアンの n-gram モデルであり、以下の式のように $(n-1)$-gram 分布が n-gram 分布を生成すると仮定している。

$$G_n \sim \mathrm{PY}(G_{n-1}, d_n, \theta_n)$$

ここで、d_n はディスカウント係数、θ_n は G_n が平均的に基底測度 G_{n-1} に似ているかを制御するパラメータである。これによりユニグラム分布からバイグラム分布が、バイグラム分布からトライグラム分布が生成される。実際に文脈 h が与えられたときの n-gram 確率は、文脈 h が与えられたときの w の出現回数 $p(w|h)$ と単語 w が 1 つ短い文脈 h' から生成されたと推定された回数 t_{hw} を用いて以下の式から再帰的に計算される。

$$p(w|h) = \frac{c(w|h) - d \cdot t_{hw}}{\theta + c(h)} + \frac{\theta + d \cdot t_{h\cdot}}{\theta + c(h)} p(w|h')$$

ここで、$t_{h\cdot} = \Sigma_w t_{hw},\ c(h) = c(w|h)$ とした。

NPYLM では部分文字列の単語らしさを推定するために、HPYLM を ∞-gram モデルに拡張した VPYLM [40] を用いて文字モデルを構築し、単語モデルの HPYLM のゼログラム基底測度に文字モデルの VPYLM をネストする。学習の際には MCMC 法を使い、前向き後向きアルゴリズムにより確率的に単語分割を行う。最大単語長を L としたとき、部分文字列 $c_1 c_2 \cdots c_t = c_1^t$ の最後の k 文字を単語として生成する前向き確率 $\alpha[t][k]$ は以下の式で表され、文末まで再帰的に計算していく。

$$\alpha[t][k] = \sum_{j=1}^{L} p(c_{t-k+1}^{t} | c_{t-k-j+1}^{t-k}) \alpha[t-k][j] \qquad (1)$$

単語分割位置のサンプリングは、文末から後ろ向きに求まる。文末の特別な記号を EOS とし入力文字列の長さを N とすると、$p(\mathrm{EOS}|c_{N-k}^{N})\alpha[N][k]$ に比例する確率で k をサンプリングし、文字列の最後の単語が得られる。その前の単語も今決めた単語に前接するようにサンプリングでき、これを文字列の先頭まで繰り返すことで入力文字列から単語列をサンプリングできる。サンプリングした単語列を通常の HPYLM と同様に MCMC の中で削除と再追加を繰り返すと言語モデルを最適化するようなパラメータが得られる。

NPYCRF [38] は、分かち書きがラベル付けされた言語資源の基準に従う単語分割を行うために、JESS-CM [41] に基づき生成モデルである NPYLM と識別モデルである CRF を統合した半教師あり学習のモデルである。JESS-CM は、入力文字列を $\boldsymbol{x}$、そのラベル列を $\boldsymbol{y}$ としたとき以下の数式で表される。

$$p(\boldsymbol{y}|\boldsymbol{x}) \propto p_{\mathrm{DISC}}(\boldsymbol{y}|\boldsymbol{x};\Lambda)\ p_{\mathrm{GEN}}(\boldsymbol{y},\boldsymbol{x};\Theta)^{\lambda_0} \qquad (2)$$

ここで、p_{DISC} は識別モデル、p_{GEN} は生成モデルであり、Λ, Θ はそれぞれのパラメータである。p_{DISC} を CRF のような対数線形モデルとすると以下の式のように表すことができ、$p(\boldsymbol{y}|\boldsymbol{x})$ は次のようになる。

$$p_{\mathrm{DISC}}(\boldsymbol{y}|\boldsymbol{x}) \propto \exp\left[\sum_{m=1}^{M} \lambda_m f_M(\boldsymbol{y},\boldsymbol{x})\right]$$

$$p(\boldsymbol{y}|\boldsymbol{x}) \propto \exp\Bigg[\lambda_0 \log(p_{\mathrm{GEN}}(\boldsymbol{y},\boldsymbol{x})) + \sum_{m=1}^{M} \lambda_m f_m(\boldsymbol{y},\boldsymbol{x})\Bigg]$$

$$= \exp(\Lambda^* \cdot F(\boldsymbol{y},\boldsymbol{x}))$$

ここで

$$\Lambda^* = (\lambda_0, \lambda_1, ..., \lambda_M)$$

$$F(\boldsymbol{y},\boldsymbol{x}) = (\log(p_{\mathrm{GEN}}(\boldsymbol{y},\boldsymbol{x})), f_1(\boldsymbol{y},\boldsymbol{x}), ..., f_M(\boldsymbol{y},\boldsymbol{x}))$$

とすると対数線形モデルで表せる。よってラベル付きデータを $\langle \boldsymbol{X_l}, \boldsymbol{Y_l} \rangle$、ラベルなしのデータを $\boldsymbol{X_u}$ とすると、

$$p(\boldsymbol{Y_l}|\boldsymbol{X_l}) = p(\boldsymbol{Y_l}|\boldsymbol{X_l};\Lambda^*)\ p(\boldsymbol{X_u};\Theta) \qquad (3)$$

が目的関数となる。学習は
(1) Θ を固定し、$\langle \boldsymbol{X_l}, \boldsymbol{Y_l} \rangle$ を用いて Λ^* を最適化
(2) Λ^* を固定し、$\boldsymbol{X_u}$ を用いて Θ を最適化
を交互に繰り返していく。CRF の情報を取り入れた NPYLM の前向き確率は式 (1) の代わりに

$$\alpha[t][k] = \sum_{j=1}^{L} \exp\Big[\lambda_0 p(c_{t-k+1}^{t} | c_{t-k-j+1}^{t-k}) + \gamma[t-k+1, t+1)\Big] \alpha[t-k][j]$$

と計算できる。ここで $\gamma[a,b)$ は a で始まり $b-1$ で終わる単語のポテンシャルであり、その経路の重みを足し合わせた値になる。また、CRF の学習 (Λ^* の学習) には、部分ラベル列の周辺確率が必要になるが、semi-Markov モデルの NPYLM と Markov モデルの CRF を統合したモデルでは

単純に計算できないため、文献 [38] では、ラベルの組み合わせごとに場合分けを行い計算している。

PYHSMM [16] は NPYLM を拡張し、教師なしで単語列とその品詞列を推定する手法である。これは単語の潜在クラスを品詞とみなし、品詞 n-gram と品詞ごとの単語生起確率を計算することで単語とその品詞をサンプリングする。文脈 h_{wz} が与えられたとき、単語 w とその品詞 z は以下の式ように計算できると仮定している。

$$p(w, z|h_{wz}) = p(w|h_w, z)p(z|h_z)$$

ここで、h_w を単語の文脈、h_z を品詞の文脈とした。$p(w|h_w, z)$ は単語列 h_w と品詞 z が与えられたときの単語 w の出現確率であり、$p(z|h_z)$ は品詞列 h_z が与えられたときの品詞 z の出現確率である。品詞集合を Z とすると、部分文字列 $c_1c_2\cdots c_t = c_1^t$ の最後の k 文字を品詞 z の単語として生成する前向き確率は以下の式で計算でき、

$$\alpha[t][k][z] = \sum_{j=1}^{L}\sum_{r=1}^{Z}\Big[p(c_{t-k+1}^{t}|c_{t-k-j+1}^{t-k}, z) \\ p(z|r)\alpha[t-k][j][r]\Big]$$

NPYLM と同様に文末から単語とその品詞をサンプリングしていく。

4. 提案手法

提案手法である教師なし NER について述べる。提案手法の概要を図 1 に示す。本手法は入力されたテキストにラベリングを行っており、単語列に対し固有表現クラスもしくはいずれの固有表現クラスにも属しないクラスである O クラスを付与する。本節では、3 節で述べた PYHSMM と semi-Markov CRF [42] を統合するモデルについて述べ、その後、そのモデルを教師なし NER に適用する手法について説明する。

4.1 PYHSMM と semi-Markov CRF の統合

文献 [38] は JESS-CM を用いて semi-Markov モデルの NPYLM と Markov モデルの CRF という異なる構造を持つモデルを統合する手法を提案しているが、これは単語分割において、カタカナ語などの単語長が大きい単語は semi-Markov CRF ではうまく分割できず、また計算量も膨大になることからである。しかし、NER など 4~5 単語の連結で十分な場合では Markov CRF よりも semi-Markov CRF が高精度な解析を行えることもある [42]。そこで我々はまず、PYHSMM と semi-Markov CRF を JESS-CM に基づき統合するモデルを提案する。

Semi-Markov CRF では入力列 $\boldsymbol{x} = (x_1x_2...x_N)$ に対し最適な $\boldsymbol{y} = (y_1y_2...y_P)$ を推定する。この y_i は (z_i, b_i, e_i) の 3 つ組から構成され、z_i, b_i, e_i はそれぞれ y_i のラベル、y_i の開始位置、y_i の終了位置である。semi-Markov CRF は以下の式で表される。

$$p(\boldsymbol{y}|\boldsymbol{x}, \Lambda) = \frac{\exp(\Lambda \cdot F(\boldsymbol{x}, \boldsymbol{y}))}{Z(\boldsymbol{x})}$$
$$Z(\boldsymbol{x}) = \sum_{\boldsymbol{y'}} \exp(\Lambda \cdot F(\boldsymbol{x}, \boldsymbol{y'}))$$

ここで $F(\boldsymbol{x}, \boldsymbol{y})$ は素性関数であり、y_i については $F(z_i, z_{i-1}, \boldsymbol{x}, b_i, e_i)$ と書くことができる。semi-Markov CRF の前向き確率は以下の式で計算できる。

$$\alpha_{\mathrm{sCRF}}[t][k][z] = \sum_{j=1}^{L}\sum_{r=1}^{Z}\Big[\alpha_{\mathrm{sCRF}}[t-k][j][r] \\ \exp(\Lambda \cdot F(z, r, \boldsymbol{x}, t-k-j+1, t-k))\Big]$$

この semi Markov CRF の前向き確率を用いて PYHSMM の前向き確率を定義すると、

$$\alpha[t][k][z] = \sum_{j=1}^{L}\sum_{r=1}^{Z}\exp\Big[\lambda_0 p(c_{t-k+1}^{t}|c_{t-k-j+1}^{t-k}, z)p(z|r) \\ + \gamma(t-k+1, t, z, r)\Big]\alpha[t-k][j][r]$$

$$\gamma(e_i, b_i, z_i, z_{i-1}) = \Lambda \cdot F(z_i, z_{i-1}, \boldsymbol{x}, b_i, e_i)$$

となり、この PYHSMM を式 (2) の生成モデル p_{GEN} とし、semi-Markov CRF を識別モデル p_{DISC} とすることで、PYHSMM と semi-Markov CRF を統合する。その後、NPYCRF と同様に式 (3) の Λ^* と Θ を交互に最適化していく。

4.2 分野特有の教師なし NER

4.1 節で提案したモデルを用いて分野特有の教師なし NER を行う。まず、PYHSMM の分割を単語単位に変更する。単語 HPYLM をチャンク HPYLM に変更し、文字 VPYLM を単語 VPYLM に変更する。その後、少数のシードデータを semi-Markov CRF に与えることで、シードデータに似た文脈で現れる単語列をチャンクと推定できる。しかし、少数のシードデータと大量のラベルなしデータを用いて、単純にチャンクの分割位置を推定すると、高頻度で出現する名詞・助詞などの単語列をまとめあげ、それをチャンクと推定する。例えば、コーパス中に「彼-が」や「彼女-は」などの単語列が多数出現するとそれらをチャンクと推定するが、これらはチャンクではない。そこで一般分野のコーパスを別に用意する。一般分野のコーパスの単語列を擬似教師データとして用いる。この擬似教師データでは、一般分野のコーパスはすべて 1 単語ごとに分割されており、かついずれの固有表現でもないクラスである O クラスに属するとする。この擬似教師データは、分野特有の固有表

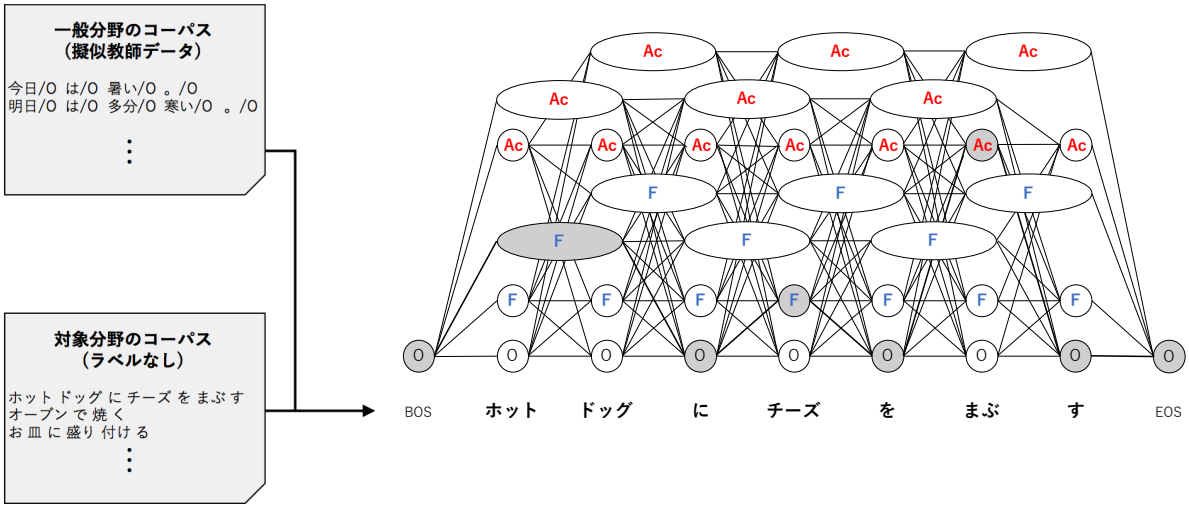

図 1 提案手法の概要 (F=食材名クラス、Ac=調理者の動作名クラス、O=固有表現ではないクラス)

Add initial segmentation $\langle \boldsymbol{X_l}, \boldsymbol{Y_l} \rangle$ to Θ
Add initial segmentation $\langle \boldsymbol{X_{pl}}, \boldsymbol{Y_{pl}} \rangle$ to Θ
Add initial segmentation $\boldsymbol{l}(\boldsymbol{X_u})$ to Θ
Optimize Λ^* on $\langle \boldsymbol{X_l}, \boldsymbol{Y_l} \rangle$
for $j = 1, 2, ..., J$ **do**
  for s =randperm($\boldsymbol{X_u}$, $\boldsymbol{X_{pl}}$) **do**
    if $j > 1$ **then**
      Remove customers of $\boldsymbol{l}(s)$ from Θ
    end if
    if $s \in X_u$ **then**
      Sample $\boldsymbol{l}(s)$ according to $p(\boldsymbol{l}|s, \Lambda^*, \Theta)$
    else
      Determine $\boldsymbol{l}(s)$ according to $\langle s, \boldsymbol{Y_{pl}} \rangle$
    end if
    Add customers of $\boldsymbol{l}(s)$ to Θ
  end for
  Optimize Λ^* on $\langle \boldsymbol{X_l}, \boldsymbol{Y_l} \rangle$
end for

図 2 分野特有の教師なし NER の学習アルゴリズム

現を含んでいる可能性があるため、そのまま semi-Markov CRF の学習データとして用いることはできない。この擬似教師データを単語単位の PYHSMM にのみ与え、弱い制約として扱う。実際の単語単位の PYHSMM の学習において、擬似教師データでは MCMC の削除と再追加を行う際にパラメータをサンプリングするのではなく、固定すれば弱い制約として扱える。また O クラスには 1 単語からなるチャンクのみが属することができる制約を与えることで、対象となる分野にのみ高頻度で出現する単語列に対して、シードデータに即した分類ができる。図 2 に分野特有の教師なし NER の学習アルゴリズムを示す。

文献 [35] の手法では、与えられたエンティティを少数のシードデータから分類することができたが、提案手法は特定分野の文書に対し、少数のシードデータのみからエンティティの検出と分類を同時にかつテキスト上で行える。例えば、GENIA コーパスに出現する「PML」は文脈により、protein 名クラスや DNA 名クラス、cell type 名クラス、O クラスに分類できる。同様に、レシピコーパスに出現する「水」は文脈により、食材名クラスや道具名クラス、O クラスに分類できる。本手法では、与えられたテキスト上から適切なクラスを推定できる。

また、文献 [36] の手法はシードデータと NP chunker を用いて名詞句を抽出し、それらを用語辞典を用いてクラス分類をおこなうため、名詞句のみの分類になってしまう。しかし、提案手法は分野特有の動詞などにも対応することができ、特別な用語辞典は必要としないため、様々な分野での応用が期待できる。

5. 実験

提案した手法を評価するために実験を行った。

5.1 対象分野とコーパス

実験に用いたコーパスの諸元を表 1 に示す。実験では英語文書のバイオ医療 NER と日本語文書のレシピ NER とゲーム解説 NER を行った。

バイオ医療のコーパスには GENIA コーパス [5] を用いて、BioNLP/NLPBA 2004 shared task [43] のテストコーパスで評価した。このコーパスでは、分野特有の固有表現クラスとして protein 名クラスや DNA 名クラスなどが 5 種類のクラスが定義されている。英語文書の一般分野コーパスには Brown コーパス [44] を用いた。これは 15 分野からテキストを集めた書き言葉コーパスである。

レシピコーパス [6] は料理レシピのテキストであり、分野特有の固有表現クラスとして食材名クラスや道具名クラス、調理者の動作名クラスなど 8 種類が定義されている。ゲーム解説コーパス [45] は将棋の解説文からなり、分野特有の固有表現クラスとして人名クラスや駒名クラス、プレ

言語	データセット	文数	単語数	固有表現数	固有表現クラス数
英語	対象分野				
	GENIA コーパス				
	学習 (ラベルなし)	10,000	264,743	-	-
	テスト	3,856	101,039	90,309	5
	一般分野				
	Brown コーパス	50,000	1039,886	-	-
日本語	対象分野				
	レシピコーパス				
	学習 (ラベルなし)	10,000	244,648	-	-
	テスト	148	2,667	869	8
	ゲーム解説コーパス				
	学習 (ラベルなし)	10,000	398,947	-	-
	テスト	491	7,161	2,365	21
	一般分野				
	BCCWJ	40,000	936,498	-	-
	話し言葉	10,000	124,031	-	-

表 1 コーパス諸元

features
$chunk_{i-1}$, $chunk_i(w_{b_i}w_{b_1+1}...w_{e_i})$, $chunk_{i+1}$
w_{b_i}, w_{b_i+1}, ..., w_{e_i}

表 2 Semi-Markov CRF に用いた素性

イヤーの動作名クラスなど 21 種類が定義されている。これら 2 つのコーパスの NE は名詞句以外からも構成されており、調理者の動作名クラスやプレイヤーの動作名クラスは動詞からなる。日本語の一般分野のコーパスには現代日本語書き言葉均衡コーパス (BCCWJ) [46] と話し言葉コーパス [47] を用いた。BCCWJ は複数の分野からなる書き言葉のコーパスである。話し言葉のコーパスとして会話作文英語表現辞典から例文の日本語を抽出し用いた。日本語のコーパスにおいて分かち書きがラベル付けされていない文には KyTea [48] *1 を用いて単語分割を行った。

5.2 実験設定

PYHSMM では、パラメータの初期値として 1 文全体を単語とみなし、そこから最適な単語列を推定する。本実験では、ラベルなしデータの初期値として 1 文全体をチャンクとし、そのクラスを O クラスとした。また、PYHSMM では日本語単語分割において、文字種ごとに異なる最大単語長を設定していたが、本実験では NE インスタンスの最大単語長を $L = 6$ とし、文字種ごとの設定はしなかった。PYHSMM に対する重み λ_0 と semi-Markov CRF の重み $\lambda_1, \lambda_2, ..., \lambda_M$ の初期値にはそれぞれガウス分布 $N(\mu, \sigma^2)$ から与え、本実験では $\mu = 1.0, \sigma = 1.0$ に設定した。semi-Markov CRF の正則化項には L2 正則化項を採用し、正則パラメータ C は 1.0 に設定し、確率的勾配降下法 (SGD) により最適化を行った。表 2 に実験で用いた semi-Markov CRF の素性を示す。現在注目している単語列 $chunk_i$ は単語 n-gram から構成されており ($\boldsymbol{w}_{b_i}^{e_i} = w_{b_i}w_{b_i+1}...w_{e_i}$)、$chunk_{i-1}$ と $chunk_{i+1}$ は前接・後接する最大長の chunk である。

本実験に使用した NER コーパスはすべて、学習用とテスト用が公開されている。学習用コーパスに最も多く出現した上位 2 つの NE インスタンスをクラス毎に抽出し、それら NE インスタンスを含むラベル付きテキストをランダムに 1 文ずつ抽出し、シードデータとして用いた。例えば、GENIA コーパスの DNA 名クラスでは、“IL-2” と “LTR” が最も多く出現し、それらを含む “IL-2 gene expression and NF-kappa B activation through CD28 requires reactive oxygen production by 5-lipoxygenase.” と “The formation of this complex would increase, independently of an in synergy with NF-kappa B, the low basal activity of the HIV LTR observed in normal T lymphocytes.” をシードデータとして与えた。

5.3 実験結果

GENIA コーパスに関しては、医療用語辞書を用いた辞書マッチングベースの手法 [49] (MetaMap) とバイオ医療の教師なし NER [36] (Zhang and Elhadad) と比較した。レシピコーパスとゲーム解説コーパスに関しては、教師なし手法による NER の従来研究がないため、ツール *2 を用いて専門用語抽出を行い、その後、抽出した専門用語に対してベイジアン HMM [50] を用いてクラス分類を行ったものベースラインとした。

*1 http://www.phontron.com/kytea/

*2 http://gensen.dl.itc.u-tokyo.ac.jp/termextract.html

対象分野 Method	Prec.	Recall	F-meas.
GENIA			
MetaMap [49]	N/A	N/A	7.70
Zhang and Elhadad [36]	**15.40**%	15.00%	15.20
Proposed	11.20%	**29.10**%	**16.30**
レシピ			
Baseline	**49.78**%	25.89%	34.07
Proposed	36.45%	**42.58**%	**39.28**
ゲーム解説			
Baseline	52.75%	29.18%	37.57
Proposed	**75.57**%	**35.05**%	**47.89**

表 3 実験結果 (Precision Recall F-measure)

実験結果を表 3 に示す。表より、すべての分野において提案手法の方が従来手法よりも精度がよいことがわかる。

6. おわりに

本稿では、特定分野における教師なし NER を提案した。本手法は PYHSMM を拡張したベイジアンモデルである。一般分野の文書を擬似教師データとして扱うことで、特定分野に固有な表現を認識・クラス分類を同時に行う。

今後の展望としては、エンティティが与えられた状態での分類性能を調査し、文献 [35] との比較を行いたい。現在は少ない素性を用いているが、品詞情報なども重要な素性と考えることができるので、素性を変更しての追加実験を行いたい。分野特有の NER の精度評価だけでなく、関係抽出や情報検索などの後段の処理において、本手法を適用した際の評価なども行いたい。また、本手法は対象となる分野にのみ高頻度で出現する単語列を NE インスタンスとして認識できるが、低頻度の NE インスタンスは認識するのが難しい。低頻度だが重要な NE インスタンスの認識を今後の課題としたい。

参考文献

[1] Thompson, P. and Dozier, C. C.: Name Searching and Information Retrieval, *CoRR*, Vol. cmp-lg/9706017 (1997).
[2] Lao, N. and Cohen, W. W.: Relational retrieval using a combination of path-constrained random walks, *Machine Learning*, Vol. 81, No. 1, pp. 53–67 (2010).
[3] Feldman, R. and Rosenfeld, B.: Boosting Unsupervised Relation Extraction by Using NER, *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pp. 473–481 (2006).
[4] Lee, H., Recasens, M., Chang, A., Surdeanu, M. and Jurafsky, D.: Joint Entity and Event Coreference Resolution across Documents, *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 489–500 (2012).
[5] Kim, J.-D., Ohta, T., Tateisi, Y. and Tsujii, J.: GENIA corpusA semantically annotated corpus for bio-textmining, Vol. 19 Suppl 1, pp. i180–2 (2003).
[6] Mori, S., Maeta, H., Yamakata, Y. and Sasada, T.: Flow Graph Corpus from Recipe Texts, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pp. 2370–2377 (2014).
[7] Yamakata, Y., Imahori, S., Maeta, H. and Mori, S.: A method for extracting major workflow composed of ingredients, tools, and actions from cooking procedural text, *2016 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, pp. 1–6 (2016).
[8] Rosario, B. and Hearst, M. A.: Classifying Semantic Relations in Bioscience Texts, *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics* (2004).
[9] Uzuner, Ö., South, B. R., Shen, S. and DuVall, S. L.: 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text, *Journal of the American Medical Informatics Association*, Vol. 18, No. 5, pp. 552–556 (2011).
[10] Islamaj Dogan, R. and Lu, Z.: An improved corpus of disease mentions in PubMed citations, *BioNLP: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*, pp. 91–99 (2012).
[11] Doğan, R. I., Leaman, R. and Lu, Z.: NCBI disease corpus: a resource for disease name recognition and concept normalization, *Journal of biomedical informatics*, Vol. 47, pp. 1–10 (2014).
[12] Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K. and Dyer, C.: Neural Architectures for Named Entity Recognition, *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 260–270 (2016).
[13] Ma, X. and Hovy, E.: End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1064–1074 (2016).
[14] Tang, S., Zhang, N., Zhang, J., Wu, F. and Zhuang, Y.: NITE: A Neural Inductive Teaching Framework for Domain Specific NER, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2642–2647 (2017).
[15] Settles, B.: Biomedical Named Entity Recognition Using Conditional Random Fields and Rich Feature Sets, *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, pp. 33–38 (2004).
[16] Uchiumi, K., Tsukahara, H. and Mochihashi, D.: Inducing Word and Part-of-Speech with Pitman-Yor Hidden Semi-Markov Models, *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1774–1782 (2015).
[17] Augenstein, I., Das, M., Riedel, S., Vikraman, L. and McCallum, A.: SemEval 2017 Task 10: ScienceIE - Extracting Keyphrases and Relations from Scientific Publications, *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pp. 546–555 (2017).
[18] Price, P. J.: Evaluation of spoken language systems: The ATIS domain, *Speech and Natural Language: Proceedings of a Workshop*, pp. 91–95 (1990).
[19] Tjong Kim Sang, E. F. and De Meulder, F.: Introduction to the CoNLL-2003 Shared Task: Language-independent Named Entity Recognition, *Proceedings of*

the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4, CoNLL '03, pp. 142–147 (2003).

[20] 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).

[21] Sang, E. F. T. K.: Introduction to the CoNLL-2002 shared task: language-independent named entity recognition, proceedings of the 6th conference on Natural language learning, *August*, Vol. 31, pp. 1–4 (2002).

[22] Grishman, R. and Sundheim, B.: Message understanding conference-6: A brief history, *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*, Vol. 1 (1996).

[23] Sekine, S. and Isahara, H.: IREX: IR and IE Evaluation project in Japanese, *Proceedings of International Conference on Language Resources & Evaluation* (2000).

[24] Bikel, D. M., Miller, S., Schwartz, R. and Weischedel, R.: Nymble: a high-performance learning name-finder, *Proceedings of the fifth conference on Applied natural language processing*, pp. 194–201 (1997).

[25] Asahara, M. and Matsumoto, Y.: Japanese named entity extraction with redundant morphological analysis, *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pp. 8–15 (2003).

[26] Borthwick, A. E.: A Maximum Entropy Approach to Named Entity Recognition, PhD Thesis (1999). AAI9945252.

[27] Lafferty, J. D., McCallum, A. and Pereira, F. C. N.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pp. 282–289 (2001).

[28] McCallum, A. and Li, W.: Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons, *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pp. 188–191 (2003).

[29] Yadav, S., Ekbal, A., Saha, S. and Bhattacharyya, P.: Entity Extraction in Biomedical Corpora: An Approach to Evaluate Word Embedding Features with PSO based Feature Selection, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, Association for Computational Linguistics, pp. 1159–1170 (2017).

[30] Ciaramita, M., Gangemi, A., Ratsch, E., Šaric, J. and Rojas, I.: Unsupervised Learning of Semantic Relations Between Concepts of a Molecular Biology Ontology, *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, pp. 659–664 (2005).

[31] Bick, E.: A Named Entity recognizer for Danish, *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)* (2004).

[32] Rau, L. F.: Extracting Company Names from Text, *Proceedings of the Seventh Conference on Artificial Intelligence Applications CAIA-91 (Volume II: Visuals)*, pp. 189–194 (1991).

[33] Sekine, S. and Nobata, C.: Definition, Dictionaries and Tagger for Extended Named Entity Hierarchy., *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)* (2004).

[34] Riloff, E. and Jones, R.: Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping, *Proceedings of the Sixteenth National Conference on Artificial Intelligence and Eleventh Conference on Innovative Applications of Artificial Intelligence*, pp. 474–479 (1999).

[35] Collins, M. and Singer, Y.: Unsupervised models for named entity classification, *Proceedings of the Joint SIGDAT Conference on Empirical Methods in NLP and Very Large Corpora* (1999).

[36] Zhang, S. and Elhadad, N.: Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts., *Journal of Biomedical Informatics*, Vol. 46, No. 6, pp. 1088–1098 (2013).

[37] Mochihashi, D., Yamada, T. and Ueda, N.: Bayesian Unsupervised Word Segmentation with Nested Pitman-Yor Language Modeling, *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pp. 100–108 (2009).

[38] Fujii, R., Domoto, R. and Mochihashi, D.: Nonparametric Bayesian Semi-supervised Word Segmentation, *Transactions of the Association for Computational Linguistics*, Vol. 5, pp. 179–189 (2017).

[39] Teh, Y. W.: A Hierarchical Bayesian Language Model Based On Pitman-Yor Processes, *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pp. 985–992 (2006).

[40] Mochihashi, D. and Sumita, E.: The Infinite Markov Model, *Advances in Neural Information Processing Systems 20, Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 3-6, 2007*, pp. 1017–1024 (2007).

[41] Suzuki, J. and Isozaki, H.: Semi-Supervised Sequential Labeling and Segmentation Using Giga-Word Scale Unlabeled Data, *Proceedings of ACL-08: HLT*, Association for Computational Linguistics, pp. 665–673 (2008).

[42] Sarawagi, S. and Cohen, W. W.: Semi-Markov Conditional Random Fields for Information Extraction, *Advances in Neural Information Processing Systems 17*, pp. 1185–1192 (2005).

[43] Kim, J.-D., Ohta, T., Tsuruoka, Y., Tateisi, Y. and Collier, N.: Introduction to the Bio-Entity Recognition Task at JNLPBA, *Proceedings of the International Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA-04)*, pp. 70–75 (2004).

[44] Francis, W. N. and Kucera, H.: Brown corpus manual, *Brown University*, Vol. 2 (1979).

[45] Mori, S., Richardson, J., Ushiku, A., Sasada, T., Kameko, H. and Tsuruoka, Y.: A Japanese Chess Commentary Corpus, *Proceedings of the Tenth International Conference on Language Resources and Evaluatio*, pp. 1415–1420 (2016).

[46] Maekawa, K., Yamazaki, M., Ogiso, T., Maruyama, T., Ogura, H., Kashino, W., Koiso, H., saya Yamaguchi, M., Tanaka, M. and Den, Y.: Balanced corpus of contemporary written Japanese, *Language Resources and Evaluation*, Vol. 48, pp. 345–371 (2014).

[47] Keene, D., Hatori, H., Yamada, H. and Irabu, S.: *Japanese-English Sentence Equivalents*, Asahi Press, Electronic book edition (1992).

[48] Neubig, G., Nakata, Y. and Mori, S.: Pointwise Prediction for Robust, Adaptable Japanese Morphological Analysis, *Proceedings of the 49th Annual Meeting of*

the Association for Computational Linguistics: Human Language Technologies, pp. 529–533 (2011).

[49] Aronson, A. R.: Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program., *Proceedings of the AMIA Symposium*, p. 17 (2001).

[50] Goldwater, S. and Griffiths, T.: A fully Bayesian approach to unsupervised part-of-speech tagging, *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 744–751 (2007).