コーパスからの固有表現辞書の自動構築

Building named entity dictionary from corpora

福島 健一　鍛治 伸裕　喜連川 優
Ken'ichi Fukushima, Nobuhiro Kaji, and Masaru Kitsuregawa

東京大学生産技術研究所
Institute of Industrial Science, the University of Tokyo
{ken,kaji,kiture}@tkl.iis.u-tokyo.ac.jp

Abstract: This paper proposes a method to automatically build a dictionary of named entities from corpora that boosts the the performance of named entity recognition by augmenting feature set. The aim of the use of the dictionary is to aid to recognize unknown entities that don't appear in training data and are problematic when adapting supervised learning to real-world text. Our experimental results suggest that the performance of recognizing unknown naned entity can be improved by our proposed method.

1　はじめに

テキスト中で人名・地名などの固有表現が使われている箇所を同定する固有表現抽出は、情報抽出・質問応答などの高度なテキスト解析のために不可欠な基礎技術として認識されている。固有表現抽出タスクでは通常教師あり学習が用いられ、日本語においても様々な手法が試され着実に精度を上げてきた [9, 12, 13, 10, 8]。固有表現抽出に限らず、教師あり学習を用いた言語処理では教師データに現れない現象の扱いが問題となる。固有表現抽出に使える教師データはごく少数であり、それらのデータを使って学習したモデルを現実的なテキストに適用するときには教師データに出現しない未知固有表現の問題は避けて通れない。我々は、固有表現辞書をコーパスから自動構築し、これを補助的な知識として用いて未知固有表現の認識精度を高める手法を提案する。この辞書は人間が使うことではなく学習の際に素性として用いることが目的なので多少のノイズの混入は許容され、低コストでの構築が可能である。日本語固有表現抽出において標準的なデータセットである CRL データを使った実験では、手法の有効性を示唆する結果が得られた。

2　固有表現抽出

人名・地名・組織名・作品名などの固有名詞と日時・時刻・金額などの数値表現を総称して固有表現 (named entity) という。文書中の固有表現の出現箇所を同定しその種類（固有表現クラス）を判断することを固有表現抽出といい、自然言語処理分野では質問応答、情報抽出などのより高次のテキスト解析を行うために必要不可欠な基礎技術の一つとして認識されている。固有表現の定義は個々の事例に依存するものの、日本語固有表現抽出手法の評価・比較においては IREX ワークショップ [2] による定義とデータセットが通常用いられる。IREX は表 1 に示す 8 種の固有表現を定義している。

表 1: IREX による固有表現の定義

固有表現クラス		例
ARTIFACT	固有物名	ノーベル文学賞
LOCATION	地名	日本、北海道
ORGANIZATION	組織名	自民党
PERSON	人名	福田、小沢
DATE	日付	二月三日
TIME	時刻	午後三時
MONEY	金額	２０億円
PERCENT	割合	３５%

本節では、固有表現抽出タスクを表現する標準的な枠組である系列ラベリング問題と、その教師あり学習手法について説明し、この枠組が本質的に抱える未知固有表現の問題について論じる。

2.1　系列ラベリング問題としての定式化

固有表現抽出を系列ラベリング問題として解く場合、まず文をトークン（単語ないし文字）の単位に分解し、固有表現を構成する（一つ以上の）トークンを再度まとめあげる。まとめあげの状態と固有表現クラスを表現するために固有表現タグを文中の各トークンに付与する。図 1 に「福田首相は１１月の訪米で」という文に

トークン	固有表現タグ
福	B-PERSON
田	I-PERSON
首	O
相	O
は	O
1	B-DATE
1	I-DATE
月	I-DATE
の	O
訪	O
米	B-LOCATION
で	O

図 1: タグによる固有表現のまとめあげ

対し文字単位で固有表現タグを付与した例を示す。タグは大きく分けて B、I、O の 3 種があり、それぞれ固有表現の開始位置、2 文字目以降、外側を意味する。B タグと I タグはさらに固有表現クラスによって修飾される。つまり全部で $2 \times 8 + 1 = 17$ 種の固有表現タグがある。図 1 のタグ列は「福田」が人名、「１１月」が日付、「米」が地名であることを意味している。

かくして固有表現抽出タスクは与えられたトークン列 $\boldsymbol{x} = (x_1, \ldots, x_{|\boldsymbol{x}|})$ に対して適切なタグ列 $\hat{\boldsymbol{y}} = (\hat{y}_1, \ldots, \hat{y}_{|\boldsymbol{x}|})$ を出力する問題に帰着される。このような問題のクラスを系列ラベリング問題といい、これを解くには任意のトークン列とタグ列の組 $(\boldsymbol{x}, \boldsymbol{y})$ にスコアを与える関数 $F(\boldsymbol{x}, \boldsymbol{y})$ と、最適な $\hat{\boldsymbol{y}} = \mathrm{argmax}_{\boldsymbol{y}} F(\boldsymbol{x}, \boldsymbol{y})$ を効率的に探索するアルゴリズムがあれば良い。スコア関数 $F(\boldsymbol{x}, \boldsymbol{y})$ は通常多数の調節可能なパラメータを持ち、その値は正解のタグが付与されたされた教師データ $\{(\boldsymbol{x}_n, \boldsymbol{y}_n)\}_{n=1}^N$ を使って教師あり学習の枠組で決定される。自然言語処理において、系列ラベリング問題は固有表現抽出だけでなく品詞タグ付け、チャンキングなど多くのタスクで利用される重要なモデルであり、HMM[6]、SVM[4]、パーセプトロン [1] など様々な機械学習アルゴリズムが試されている。本研究ではそれらの学習手法の中で最高の精度を持つといわれる条件付確率場 (conditional random field; CRF)[5] を用いる。

CRF は入力トークン列 $\boldsymbol{x}$ が与えられたとき、あらゆる可能なタグ列 $\boldsymbol{y}$ に対して次の式で表される条件付確率を定義する。

$$\Pr(\boldsymbol{y}|\boldsymbol{x}; \boldsymbol{\Lambda}) = \frac{1}{Z_{\boldsymbol{x}}} \exp\left(\sum_{i=1}^{|\boldsymbol{x}|-1} \sum_k \lambda_k f_k(y_i, y_{i+1}, \boldsymbol{x}, i) \right)$$

$Z_{\boldsymbol{x}}$ は $\Pr(\boldsymbol{y}|\boldsymbol{x})$ が正しい確率であるための正規化係数である。各 $f_k(y_i, y_{i+1}, \boldsymbol{x}, i)$ は素性関数と呼ばれ、これによってトークンとタグの依存関係を表現してモデルに取り込む。$\boldsymbol{\Lambda} = \{\lambda_k\}$ はモデルのパラメータであり、各素性の重みを与える。素性関数には、引数の形さえ守れば任意の構造と値域が許されるが、典型的にはトークン部分のみを引数にとる述語関数とタグの値の組合せで表現できる範囲に限定する。また x に含まれる情報として、入力トークンの表層的な文字列だけでなく、形態素解析の結果や文字種、シソーラスの情報など、ありとあらゆる情報を利用可能である。例えば素性関数として

$$f_k = \begin{cases} 1 & \text{if } y_i = \text{B-LOCATION} \wedge x_i = \text{米} \\ 0 & \text{otherwise} \end{cases}$$

や

$$f_k = \begin{cases} 1 & \text{if } y_i = \text{O} \wedge x_i\text{の品詞} = \text{助詞} \\ 0 & \text{otherwise} \end{cases}$$

などが考えられる。これらの素性関数は、“「米」という文字は地名の先頭になり得る”、“助詞は固有表現に含まれにくい” という関係を表現することを意図して設計されている。なお、素性関数の値が 1 になることを “素性が発火する” という。固有表現抽出で用いられる典型的な素性を図 2 を使って説明する。図の枠内の 要素が i 番目の固有表現タグ（正解は O）を推定するために使われる情報である。枠内の各要素と前説で説明した 17 種の固有表現タグのすべての組合せが素性関数としてモデルに組み込まれる。パラメータの最適値 $\hat{\boldsymbol{\Lambda}}$ は教師データの尤度 $\prod_{n=1}^N \Pr(\boldsymbol{y}_n|\boldsymbol{x}_n; \boldsymbol{\Lambda})$ ないし事後確率 $\Pr(\boldsymbol{\Lambda}) \prod_{n=1}^N \Pr(\boldsymbol{y}_n|\boldsymbol{x}_n; \boldsymbol{\Lambda})$ を最大化するように推定される。この推定方法はモデルでの素性関数の期待値が、教師データでのその経験的な値と等しくなるようにパラメータをとることに等しい。教師データを使って学習されたモデルは新たなトークン列 $\boldsymbol{x}'$ を与えられるとそれに対応する最適なタグ列 $\hat{\boldsymbol{y}}' = \mathrm{argmax}_{\boldsymbol{y}} \Pr(\boldsymbol{y}|\boldsymbol{x}'; \hat{\boldsymbol{\Lambda}})$ を出力する。解の候補の数はトークン列の長さの指数関数なので、動的計画法の一種である Viterbi アルゴリズムを使って効率的な探索を行う。

2.2 未知固有表現の問題

固有表現抽出タスクに機械学習を使う動機のひとつとして、辞書あるいは教師データに現れない固有表現（未知 NE）を認識する能力に対する期待がある。固有表現辞書をあらかじめ用意してマッチングを行うというアプローチだと、辞書の作成コストが非常に高く、また十分な被覆率を保つことも難しいので未知 NE に対処する方法が別途必要になる。一方で前項で説明したようなモデルならば、辞書が不要な上に、固有表現が出現しやすい文脈を帰納的に学習することで、表層的な字面に依存せずに認識することも原理的には可能である。このような背景があり、固有表現抽出に限らず自然言語処理の様々なタスクにおいて、ルールベース・知識ベースの手法よりも機械学習による手法が有望であると考えられている。

位置	文字	字種	単語	品詞	固有表現タグ
i-2	福	OTHER	B-福田	B-名詞-固有名詞-人名-姓	B-PERSON
i-1	田	OTHER	E-福田	E-名詞-固有名詞-人名-姓	I-PERSON
i	首	OTHER	B-首相	B-名詞-一般	O
i+1	相	OTHER	E-首相	E-名詞-一般	O
i+2	は	HIRA	S-は	s-助詞-係助詞	O

図 2: 日本語固有表現抽出で用いられる標準的な素性

表 2: 既知 NE と未知 NE での再現率の差異

	再現率（事例数）				
固有表現クラス	全体	既知 NE		未知 NE	
ART	51.65	83.03	(288)	31.93	(459)
LOC	89.72	95.01	(4470)	65.56	(993)
ORG	82.46	92.92	(2537)	59.12	(1139)
PSN	90.24	95.16	(305)	81.99	(187)
DAT	93.44	95.05	(2906)	86.25	(661)
TIM	87.95	90.11	(340)	83.31	(162)
MON	92.56	98.10	(128)	89.79	(262)
PNT	91.24	88.59	(305)	95.70	(187)
（全体）	87.65	94.15		71.17	

機械学習による手法は、確かに未知 NE を認識することが可能である。しかし、その認識率は教師データに出現する固有表現（既知 NE）に比べるとかなり劣ることが明らかになっている [14]。これは文字や単語を素性として使っていることからも納得できる。機械学習を用いた既存の手法は、CRL データを使った実験で 90%近い認識精度を達成しているものの [13][8]、既知 NE と未知 NE を区別してその数字が議論されたことはない。我々は既知 NE と未知 NE に対する精度の違いについて、CRL データを使って独自に調査を行った。CRF と図 2 に示した素性の組合せを使い 5 分割交差検定によって算出した既知 NE・未知 NE 別の再現率と事例数を表 2 に示す。PERCENT 以外の全てのクラスで未知 NE に対する再現率は既知 NE に対するそれに劣っている。さらに詳しく観察すると、字面による推定が容易な数値表現と形態素解析器が出力する品詞情報（図 2 の名詞-固有名詞-人名-姓など）が強い手がかりになる PERSON では数値の乖離はあまり大きくないが、残りの 3 つのクラスでは未知語の再現率はかなり悪く、特に ARTIFACT では既知 NE の 83.03%に対して未知 NE では 31.93%しかない。また ARTIFACT では未知 NE の事例数が 459 と既知 NE の 288 より多い。

未知 NE の問題は、閉じたデータセットのなかで手法の善し悪しを議論するときよりも、そのデータセットを使って学習されたモデルを現実的なデータに適用するときにより深刻な問題となる。テキスト中で使用される固有表現はそのテキストの分野や書かれた時期（ドメイン）に強く依存し、ドメインが異なれば異なるほど全固有表現に占める未知 NE の比率は高くなる。しかし教師データ作成のコストは非常に高いので、固有表現抽出を行いたいドメイン毎に教師データを作るわけにはいかず、CRL データなどのごく限られたデータを使わざるを得ない。このような理由により未知 NE の認識能力は極めて重要であり、手法の評価・比較の際に必ず考慮すべき要素だと考えられる。

3 提案手法

人手による辞書やルールの整備を必要としないことは機械学習によるアプローチのメリットではあるものの、2.2 項での未知 NE についての議論を踏まえて、我々はそのアプローチにおいても固有表現辞書などの補助的な言語リソースは重要な役割を果たしうると主張する。ただし学習アルゴリズムは一つの手がかりのみを絶対的に信頼するのではなく複数の情報を総合的に考慮して判断を下すので、そこで使う辞書は人間のためのそれのように完璧な精度を持つ必要はなく、ノイズが混入していても構わない。これは人手によるチェックを最小限に抑えた、低コストでの辞書の自動構築を可能にする。我々はコーパスから固有表現辞書を自動構築し、それを用いて未知 NE の認識性能を向上させる手法を提案する。

以下、本研究で扱う固有表現辞書の形式、それが固有表現抽出タスクにおいて有効に作用することの直観、固有表現辞書をコーパスから自動獲得する手法、具体的に素性関数として CRF のモデルに組み込む方法について説明する。

3.1 NE 辞書の利用

例えば「昨日、エビフィレオを試した」というテストデータに対して「エビフィレオ」が商品名 (ARTIFACT) であると認識することを考えよう。もし「エビフィレオ」が未知 NE ならば、それを構成する文字、単語は手がかりにならない。前後の文脈も特に商品名を示唆するようなものではなく、この文からは「エビフィレオ」が商品名であることを示唆する手がかりを見つけられず、認識に失敗してしまう。

しかし、教師データ中に別の商品名「ビッグマック」が ARTIFACT タグと伴って出現し、さらに別の知識源から「エビフィレオ」と「ビッグマック」が共通のカテゴリ「メニュー」に属することがわかるとどうなるだろうか。「メニュー」という素性は ARTIFACT タグを導

きやすいという規則を学習することによって、テストデータ中のメニュー「エビフィレオ」にも ARTIFACT タグを付与できるようになることが期待できる。この「メニュー – ビッグマック」「メニュー – エビフィレオ」という形式の知識の集合が本研究で扱う固有表現辞書（NE 辞書）である。「メニュー」を concept、「ビッグマック」「エビフィレオ」を instance と呼ぶことにする。この NE 辞書をコーパスから自動構築し、固有表現抽出における未知 NE 認識率を高めることが我々の目標である。

3.2 コーパスからの concept–instance ペアの収集

我々は形態素解析を施したコーパスに対して、次の正規表現で記述されるパターンをマッチさせることによって concept–instance ペアの収集を行う。

$$noun「\ any+\ 」 \tag{1}$$

ただし *noun* は形態素解析器が IPA 品詞体系 [11] の “名詞-一般”、“名詞-サ変接続”、“名詞-接尾-一般” のいずれかの品詞と判断する形態素、*any* は任意の形態素である。*noun* 部分が concept となり、*any*+ 部分が instance となる。このパターンを使う根拠は、“メニュー「ビッグマック」” や “メニュー「エビフィレオ」” などという表現が頻繁に用いられるという直観である。実際このカギ括弧のパターンはペアの抽出において有効であるものの、発話を表現する目的や、単なる強調の目的でも用いられ（表 3）、またそれ以外のノイズも多い。このためペアを concept、instance の両面からフィルタリングする必要がある。まず各 concept に対して、それとペアをなす instance の長さの平均、句読点や動詞の個数の平均などを計算し、これらの数値が大きな concept を捨てる。これによって表 3 の (3)、(4)、(5) のような発話のパターンの大半を除去できる。次に、大規模なコーパスから算出した文字 n-gram の頻度統計を用いて instance をフィルタリングする。これにより表 3 の (6)、(7) のような強調のパターンを除去する。これらのフィルタリングを施しても NE 辞書の精度は 100%からは程遠いものの、有用な情報が含まれていれば、学習アルゴリズムがそれをノイズと区別して抽出してくれることに期待している。

3.3 素性設計

NE 辞書の情報は図 3 のような素性設計を通してモデルに組み込まれる。入力文のある部分文字列が NE 辞書に instance として登録されている場合、その instance とペアをなすすべての concept の印をその部分文字列につける。具体的には、部分文字列中の各文字に、その文字が instance の何文字目であるかを表すタグと concept の組を付与する。位置を表すタグの体系は固有表現のまとめあげに使うそれとは異なり、B で instance の先頭を、I で途中を、E で末尾を表す。さらに一文字のみからなる instance の場合には S という特別なタグを付与する。図 3 の例は辞書に「店–マック」、「メニュー–エビフィレオ」というペアが含まれていた場合の素性の表現である。

ただし、このような素朴なマッチングでは非常に多くの素性が固有表現と無関係な箇所で発火してしまうので、いくつかのヒューリスティクスによって素性の発火を抑制する。(1)instance の開始位置、終端位置が形態素区切りと一致していない、(2)instance の前後の形態素が名詞である、(3)instance が単一の形態素からなり、その品詞が “名詞-一般”、“名詞-サ変接続” である、などの条件を満たす場合は辞書の素性を用いない。

表 3: カギ括弧パターンの典型的な用例

タイプ		用例
固有表現	(1)	高速列車「ユーロスター」に乗って
	(2)	松竹映画「男はつらいよ」シリーズ
発話	(3)	ナレーション「とか言いつつひかりにあげるものを探すタケルだった。」
	(4)	Ｓ々木「それじゃ。夜遅く掛けてごめんね。教えてくれてありがとう」
	(5)	Ｏ泉「いえいえ」
強調	(6)	かたちやスタイル「だけ」を気にしている
	(7)	この作業の区切り「まで」は片付けよう

文字	辞書素性–メニュー	辞書素性–店
マ	O	B-店
ッ	O	I-店
ク	O	E-店
で	O	O
エ	B-メニュー	O
ビ	I-メニュー	O
フ	I-メニュー	O
ィ	I-メニュー	O
レ	I-メニュー	O
オ	E-メニュー	O
を	O	O
試	O	O
す	O	O

図 3: 辞書の情報を表現する素性

4 実験

4.1 固有表現辞書の構築

今回の実験で concept-instance ペアの抽出対象とするコーパスは、1991 年から 2005 年までの 15 年分の毎

表 5: 固有表現辞書を使った場合の認識精度

固有表現クラス	F 値（全体）		再現率（既知 NE）		再現率（未知 NE）	
ARTIFACT	60.96	+1.15	78.60	-4.43	37.28	+5.35
LOCATION	89.95	-0.52	93.92	-1.09	67.60	+2.04
ORGANIZATION	84.03	-0.73	90.73	-2.19	60.16	+1.04
PERSON	90.67	-0.55	94.18	-0.98	82.29	+0.30
DATE	94.23	-0.19	95.02	-0.03	85.30	-0.95
TIME	89.25	-1.65	86.69	-3.42	80.67	-2.64
MONEY	95.60	+0.80	96.28	-1.82	91.64	+1.85
PERCENT	93.71	-0.83	88.00	-0.59	94.21	-1.49
（全体）	88.95	-0.47	92.99	-1.16	71.95	+0.78

表 4: 固有表現辞書の統計

効率 Efficiency	0.005 (424/83718)
精度 Precision	0.144 (1504/10444)
被覆率 Coverage （type ベース）	
ARTIFACT	0.266 (144/541)
LOCATION	0.084 (118/1403)
ORGANIZATION	0.046 (69/1487)
PERSON	0.047 (88/1842)
DATE	0.012 (12/950)
TIME	0.019 (4/206)
MONEY	0.000 (0/301)
PERCENT	0.000 (0/249)
（全体）	0.061 (424/6934)
被覆率 Coverage （token ベース）	
ARTIFACT	0.319 (239/747)
LOCATION	0.133 (732/5463)
ORGANIZATION	0.053 (195/3676)
PERSON	0.074 (287/3840)
DATE	0.010 (37/3567)
TIME	0.027 (14/502)
MONEY	0.000 (0/390)
PERCENT	0.000 (0/492)
（全体）	0.080 (1504/18677)

日新聞コーパス[1]と、独自に収集したウェブページから抽出したテキストを合わせたものである。テキストの量はそれぞれ 1.7GB、9.8GB である。MeCab[2]を使って形態素解析を行い、パターン (1) で concept-instance ペアを抽出した。この段階で 68 万種、述べ 93 万個のペアがあり、concept は 1.8 万種、instance は 55 万種ある。ここから発話や強調を目的とするパターンをふるい落とし、さらに教師データで固有表現タグと一度も共起しない concept を除去すると、concept の数を約 1000 種まで減らすことができる。CRF の実行時間に係わる制約から今回はこれらのうち 50 種の concept だけを素性として利用する。50 種の concept は、それとペアをなす instance が教師データ中の固有表現をどれだけ網羅するかを基準にして選んだ。これに対してさらに n-gram 頻度統計による instance のフィルタリングを施し、その結果得られた最終的な固有表現辞書はペアが 9.7 万種、instance が 8.3 万種からなる。この辞書の詳しい統計を表 4 に示す。効率は辞書中の全 instance のうち一度でも教師データ中で発火するものの割合を、精度は instance が発火したときそれが NE タグと共起する割合を、被覆率は辞書が教師データ中の固有表現をどれだけ網羅しているかを表している。

[1]ただし、CRL データの作成元となっている 1995 年 1 月分の記事は全て除外している。

[2]http://mecab.sourceforge.net/

4.2 固有表現抽出実験

構築した辞書が、教師つき学習による固有表現抽出手法の性能にどう寄与するのかを、CRL 固有表現データを用いて検証した。CRL 固有表現データは毎日新聞 1995 年版 1,174 記事、10,718 文に対して、18,677 個の固有表現が IREX の定義に従ってタグづけされている。これを記事単位で 5 分割し交差検定を行う。CRF[3]に辞書の素性を組み込んで実験した結果を表 5 に示す。表には既知 NE と未知 NE を区別せずに計算した F 値、既知 NE と未知 NE を区別して計算した再現率を、ベースライン（辞書素性を利用しない場合）との差とともに載せた。概ね既知 NE ではベースラインよりも認識率が下がり、未知 NE では上がり、また全てのタグを合わせた F 値も下がっている。

4.3 考察

固有表現辞書を使ったことで、未知 NE の認識率は目論見どおり向上した。特に ARTIFACT での上昇幅は 5.35 ポイントと大きい。今回は CRF の実行時間の制約上、辞書の一部しか使わなかったが、すべてを使うことでさらなる認識率の向上の余地がある。しかしその一方で、ベースラインの手法では認識できていた固有表現の一部を認識できなくなってしまった。これは concept-instance ペアのフィルタリングが不十分なために、辞書素性が固有表現とは無関係なところで多く発火してしまうことが原因だと考えている。表 4 からわかるように、発火する辞書素性の大半はノイズである。望ましくない発火を抑制する工夫を取り入れてはいるが、それも完全なものではない。

[3]実装は CRF++を利用した。http://crfpp.sourceforge.net/

5 関連研究

Sumida らは我々と同様のカギ括弧パターンから出発してブートストラップ的に concept-instance ペアを収集する手法を提案し、6.5GB の HTML 文書から 83.5%の高い適合率で 4,276 種のペアを獲得している [7]。我々は辞書の品質よりも固有表現抽出の精度の向上を直接的に指向している点で Sumida らと異なっている。

固有表現抽出において自動獲得した知識を補助的に用いるアプローチに、英語を対象とした Kazama[3] らの研究がある。Kazama らは Wikipedia から我々の concept-instane ペアと類似の知識を抽出している。また日本語の固有表現抽出では、自動獲得した知識を使った研究は我々の知る限りまだ報告されていないが、補助的な知識としてシソーラスを用いると精度が上がることが報告されている [12]。

6 まとめと今後の研究課題

本論文では、日本語固有表現抽出において、自動獲得した固有表現辞書を使って未知固有表現の認識精度を向上させる手法を提案した。CRL データを用いた実験を行い、本手法の有効性を示唆する結果を得た。

今後は、未知 NE の認識率をさらに高めると同時に、既知 NE の認識率を犠牲にしない手法を追求していく予定である。具体的には、辞書が持つ全情報を学習モデルに組み込めるよう工夫するとともに、より巨大なコーパスから辞書を構築することで未知 NE 認識率の向上をはかる。既知 NE の認識率の悪化に関しては、辞書素性を発火させるかどうかの基準を人手で書き下すのではなく SVM などの分類器を学習させること、素性設計をよりきめ細かく作りこむことなどで対処可能だと考えている。

参考文献

[1] M. Collins. Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms. In *Proc. of EMNLP*, 2002.

[2] IREX 実行委員会（編）. IREX ワークショップ予稿集, 1999.

[3] J. Kazama and K. Torisawa. Exploiting Wikipedia as External Knowledge for Named Entity Recognition. In *Proc. of EMNLP-CoNLL*, 2007.

[4] T. Kudo and Y. Matsumoto. Chunking with Support Vector Machines. In *Proc. of NAACL*, pp. 1–8, 2001.

[5] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. of ICML*, pp. 282–289, 2001.

[6] C.D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*, chapter 9. MIT Press, 1999.

[7] A. Sumida, K. Torisawa, and K. Shinzato. Concept-Instance Relation Extraction from Simple Noun Sequences Using a Full-Text Search Engine. In *Proc. of Workshop on Web Content Mining with Human Language Technologies*, 2006.

[8] 山田寛康. Shift-Reduce 法に基づく日本語固有表現抽出. 情報処理学会研究報告 NL-179, pp. 13–18, 2007.

[9] 山田寛康, 工藤拓, 松本裕治. Support Vector Machine を用いた日本語固有表現抽出. 情報処理学会論文誌, Vol. 43, No. 1, pp. 44–53, 2002.

[10] 福岡健太. Semi-Markov Conditional Random Fields を用いた固有表現抽出に関する研究. Master's thesis, 奈良先端科学技術大学院大学 情報科学研究科, 2006.

[11] 浅原正幸, 松本裕治. ipadic version 2.7.0 ユーザーズマニュアル, 2003.

[12] 浅原正幸, 松本裕治. 日本語固有表現抽出におけるわかち書き問題の解決. 情報処理学会論文誌, Vol. 45, No. 5, pp. 1442–1450, 2004.

[13] 中野桂吾, 平井有三. 日本語固有表現抽出における文節情報の利用. 情報処理学会論文誌, Vol. 45, No. 3, pp. 934–941, 2004.

[14] 齋藤邦子, 鈴木潤, 今村賢治. CRF を用いたブログからの固有表現抽出. 言語処理学会第 13 回年次大会論文集, 2007.