大規模コーパスを用いた固有表現抽出手法の検討

Investigation of Named Entity Extraction Method Using Large Corpora

南 和江*,　藤井 康寿*,　土屋 雅稔†,　中川 聖一*

豊橋技術科学大学

*情報 知能工学系 / †情報メディア基盤センター

1　はじめに

人名・組織名といった語句を同定する固有表現抽出タスクは，情報検索や情報抽出の基礎技術としてのみならず，自然言語処理における構文解析や意味解析などに大きな影響を及ぼすため，重要な問題である [2, 5]. 従来は，固有表現タイプとして，人名，地名，組織名など10種類程度を考慮することが一般的である．しかし，情報抽出や質問応答の適応分野の広がりを考えると，従来のような少数の分類では不十分であり，より多数の固有表現タイプを考慮する必要がある．そのため，関根は，固有表現タイプを200種類と大幅に増やし，階層的に整理分類した「関根の拡張固有表現」(以下，**拡張固有表現**と呼ぶ) を提案している [6].

固有表現抽出は，形態素列に対するチャンキング問題，または形態素を素性とする文字列に対するチャンキング問題として定式化した上で，Support Vector Machine などの機械学習手法を適用することが一般的である [9, 7]．しかし，拡張固有表現には多数のタイプが存在するため，一般的な2値分類器を pairwise 法などにより組み合わせて解く方法は，処理時間の観点から現実的ではない．そのため，新納らは，固有表現部分を抽出する処理と，抽出されたチャンクを分類する処理との2段階に処理を分割することによって，処理速度を改善する方法を提案している [8].

本研究では，機械学習手法として CRF[3] を用いることにより，多数の拡張固有表現タグを高速に付与するモデルを学習する．その訓練コーパスとして，橋本ら [11] によって作成された固有表現タグ付き大規模コーパスを用いる．このような大規模コーパスを用いる場合には，学習時間が問題になることが多いため，オンライン学習とバッチ学習の比較を行う．

2　拡張固有表現タグ付きコーパス

従来の固有表現抽出タスクでは，固有表現のタイプとして，人名，地名，組織名など10種類程度を考慮することが一般的である．例えば，MUC プロジェクト [2] では7種類，IREX プロジェクト [5] では8種類の固有表現タイプが定義されている．しかし，様々な自然言語処理技術に応用するためには，新聞記事や百科事典などに見られる各種の概念や単語を考慮する必要があり，従来のような少数の分類は不十分である．そのため，固有表現タイプを200種類と大幅に増やした**拡張固有表現**が提案されている [6].

拡張固有表現のもう1つの特徴は，固有表現タイプ間に3階層の階層関係を設定していることである (図1)．例えば，第1階層 (28種) の固有表現タイプ「地名」は，第2階層 (103種) では「地域名」「地形名」など7種類に細分されている．さらに，第2階層の固有表現タイプ「地形名」は，第3階層 (200種) では「島名」「河川名」など7種類に細分されている．

統計的機械学習により拡張固有表現抽出を行うには，訓練データとして，拡張固有表現タグが付与されたコーパスが必要である．橋本ら [11] は，日本語書き言葉均衡コーパス [10] のモニター公開データ2009年度版に含まれる白書　書籍 WEB コーパスと毎日新聞 (1995年) を対象として，拡張固有表現タグを付与したコーパスを公開している[1]．このコーパスの概要を表1に示す．本研究では，このコーパスを訓練データとして用いて，拡張固有表現抽出を行う．

3　CRF による拡張固有表現抽出

日本語の固有表現抽出においては，(1) 新規の固有表現 (例:創設された会社) が頻繁に出現するため，全ての固有表現を網羅した辞書を作成することは現実的ではない，(2) 多くの固有表現は，既知の一般語の連続 (例:東京新聞) である，という2点の理由から，固有表現抽出と形態素解析とは独立に実行可能であるという仮定を置く．ただし，形態素よりも短い単位の固有表現が存在することに注意が必要である．例えば，IREX ワークショップにおける固有表現の定義に従うと,「訪米」という形態素の部分文字列「米」を，アメ

[1] http://riverstone.star.titech.ac.jp/taiichi/tokutei/ene/

地名	地域名	大陸地域名	オリエント，北アフリカ，ゴンドワナ大陸，バビロニア，陸半球
		国内地域名	奥羽地方，中部地方，カルナティック，ボスニア，可美
		⋮	
	地形名	山地名	富士山，間ノ岳，青崩峠，中央アルプス，木曽駒ケ岳
		島名	ラクシャドウィープ諸島，友ヶ島，大スンダ列島，西表島，沖縄諸島
		⋮	
	⋮		
⋮			

図 1: 拡張固有表現の階層構造

表 1: 拡張固有表現タグ付きコーパスの概要

	白書	書籍	Web	新聞	全体
文書数	62	81	938	8255	9336
平均文書長 (文字数)	5754.5	4524.5	383.1	452.2	2778.6
タグ総数 (コーパス全体)	11819	14206	5609	240755	272389
平均タグ出現頻度 (文字当り)	0.033	0.039	0.016	0.065	0.011

位置	表層形	形態素	品詞	文字種	固有表現タグ
i-2	茨	B-茨城	B-名詞-固有名詞	OTHER	B-LOCATION
i-1	城	E-茨城	E-名詞-固有名詞	OTHER	I-LOCATION
i	県	B-県内	B-名詞-一般	OTHER	I-LOCATION
i+1	内	E-県内	E-名詞-一般	OTHER	O
i+2	の	S-の	S-助詞-連帯化	HIRA	O

図 2: 素性

表 2: 拡張固有表現タイプと出現頻度

	第 1 階層のみ区別	第 3 階層まで区別
100 回以上出現したタイプ	24	145
10 回以上出現したタイプ	1	40
1 回以上出現したタイプ	3	15
出現しなかったタイプ	0	0

リカを意味する固有表現として抽出しなければならない．そのため，浅原ら [9] は，固有表現抽出を，形態素情報を素性とする文字列に対するチャンキング問題として定式化している．

本研究でも，浅原らの定式化に従い，拡張固有表現抽出を，図 2 のように形態素情報と文字種情報を素性とする文字列に対するチャンキング問題として定式化する．さらに，タグ列の決定には CRF[3] を用いることにする．CRF では，文字列 X に対するタグ列 Y の条件付き確率 $P(Y|X)$ を，次式のように表す．

$$P(Y|X) = \frac{1}{Z(X)} \exp\left(\sum_i^n \sum_k \lambda_k f_k(X_i, Y_i)\right), \quad (1)$$

ここで，f_k は素性関数，λ_k は素性関数に対する重み，$Z(X)$ は正規化項である．

重み $\boldsymbol{\lambda}$ は，L-BFGS[4] を用いて，全訓練事例を用いた繰り返し学習によって求めることが一般的である．しかし，このようなバッチ学習手法は，訓練コーパスの分量が増加すると，それ以上に学習時間が増えてしまう問題がある．そのため，大規模な訓練コーパスに対しても適用できる学習手法として，確率的勾配降下法 (Stochastic Gradient Descent; SGD) などのオンライン学習手法が注目を集めている．本研究では，FOBOS[1] という各種の正則化に対応した SGD を用いる．FOBOS では，重み $\boldsymbol{\lambda}$ を以下の式により更新する．

$$\boldsymbol{\lambda}_{t+\frac{1}{2}} = \boldsymbol{\lambda}_t - \ {}_t\boldsymbol{g}_t^b \quad (2)$$

$$\boldsymbol{\lambda}_{t+1} = \operatorname*{argmin}_{\boldsymbol{\lambda}} \left\{ \frac{1}{2} ||\boldsymbol{\lambda} - \boldsymbol{\lambda}_{t+\frac{1}{2}}||^2 + \ {}_t r(\boldsymbol{\lambda}) \right\} \quad (3)$$

ここで，$\boldsymbol{g}_t^b$ は訓練コーパスの一部から求めた勾配，$r(\boldsymbol{\lambda})$ は正則化項，${}_t$ は学習係数である．

4 実験

拡張固有表現タイプは，第 3 階層まで全ての階層を区別すると 200 種類になる．橋本らの作成したコーパスは，IREX プロジェクトによる既存の日本語固有表現タグ付きコーパス (毎日新聞 1995 年 1 月 1 日 ∼1 月 10 日) と比較すると，かなり大規模なコーパスである．それでも，第 3 階層まで全てを区別すると，表 2 より，15 種類の拡張固有表現タイプについては抽出規則を学習することは非常に困難と予想される．そのため，本研究では，第 1 階層までを区別し，28 種類の拡張固有表現を抽出するというタスク設定で実験を行った．

抽出性能と学習時間　最初に，オンライン学習アルゴリズム (FOBOS) とバッチ学習アルゴリズム (L-BFGS) について，学習に要する時間の比較を行った．結果を表 3 に示す．この実験は，サブコーパス毎に 5 分割してから，新聞サブコーパス 1/5，白書サブコーパス 1/5，書籍サブコーパス 1/5，WEB サブコーパス 1/5

を結合してテストコーパスとし，残りを訓練コーパスとする試行を 5 回繰り返す 5 分割交差検定で実施した．表 3 より，コーパス全体を用いて実験した場合には，FOBOS は，L-BFGS と比較して約 5 倍速いことが分かる．ただし，$F_{\beta=1}$ 値でみると，L-BFGS が，FOBOS よりも 1.9 ポイント優れている[2]．さらに，L-BFGS について，訓練時間と性能の関係を見るために，訓練コーパスの量を $1/2, 1/4, 1/10$ とした実験を行った．L-BFGS と FOBOS の訓練時間が最も近づいている条件は，訓練コーパスを 1/4 に減らした場合だが，この場合は L-BFGS は FOBOS に比べてかなり性能が劣化している．以上より，拡張固有表現タイプを全て十分に含むような大規模コーパスを対象とする場合には，FOBOS などのオンライン学習アルゴリズムを適用する必要があると考えられる．そのため，以下の検討は，全て FOBOS を用いて行った．

次に，拡張固有表現タイプ別の抽出性能を表 4 に示す．表 4 より，訓練コーパス中の出現頻度と抽出性能には，かなり関連があることが分かる．ただし，例外的な拡張固有表現タイプとして，イベント名と自然物名は抽出が比較的困難であり，逆に倍数表現は容易であることが分かる．

訓練コーパスの分量　続いて，サブコーパス毎に訓練コーパスとして分量が十分かどうかの検討を行った．まず，新聞サブコーパスを均等に 10 ブロックに分割し，1 ブロックをテストコーパスとして取り除いておく．残りの 9 ブロックを訓練コーパスとして使い，訓練コーパスの分量と抽出性能との関係を調べた結果を図 3 に示す．図 3 より，訓練コーパスの増加によって，既知固有表現率[3]は単調に改善しているが，抽出性能の改善については頭打ちの傾向が見られる．よって，より大規模なコーパスを用いて実験を行わないと断定できないが，新聞サブコーパスは訓練コーパスとして十分な分量を備えているように思われる．

同様の実験を，他の 3 つのサブコーパスに対して行った結果を図 4~6 に示す．いずれのサブコーパスについても，抽出性能は，新聞サブコーパスに比べてかなり低い．また，既知固有表現率も新聞サブコーパスよりかなり低いことから，白書，書籍，WEB の 3 つのサブコーパスについては，サブコーパス単体では訓

[2]IREX ワークショップの定義に基づく固有表現抽出では，$F_{\beta=1}$ 値として 89.9 が報告されている [7]．分類が 8 種類から 28 種類に増えて，タスクが困難になっていることを考慮すると，FOBOS,L-BFGS ともに先行研究と同等以上の性能を達成していると言える．

[3]既知固有表現率は，以下の比率である．

$$\frac{\text{テストコーパスと学習コーパスの両方に出現した固有表現}}{\text{テストコーパスに出現した固有表現}}$$

表 3: 抽出性能および学習時間の比較

アルゴリズム	コーパス量	Rec.	Pre.	$F_{\beta=1}$	所要時間
FOBOS	1/1	88.6	90.0	89.3	16h
L-BFGS	1/1	90.1	92.2	91.2	79h
	1/2	83.1	86.5	84.8	41h
	1/4	78.0	83.1	80.5	21h
	1/10	71.4	79.0	75.0	7h

表 4: 拡張固有表現タイプ別の抽出性能

	Rec.	Pre.	$F_{\beta=1}$	頻度
製品名	83.7	88.6	86.0	66801
地名	91.8	91.6	91.7	39185
人名	94.6	94.2	94.4	32400
組織名	87.4	88.5	87.9	30990
時間	94.3	94.1	94.2	25093
個数	89.7	92.5	91.1	15400
施設名	84.9	83.4	84.1	10924
イベント名	80.2	78.0	78.4	8768
自然物名	76.3	79.2	77.6	7756
期間	90.1	91.2	90.7	6043
年齢	96.3	96.9	96.6	4438
金額表現	95.6	96.7	96.1	4264
割合表現	95.0	95.5	95.2	3724
序数	88.4	89.3	88.9	3600
寸法表現	92.1	91.2	91.6	3049
順位表現	83.9	89.3	86.3	2124
ポイント	83.1	84.5	83.6	1631
病気名	83.8	92.6	87.7	1544
学齢	88.1	91.1	89.6	1010
数値表現_その他	74.3	73.2	72.2	730
色名	76.0	83.8	79.5	551
頻度表現	67.1	66.5	65.9	377
倍数表現	94.5	95.9	95.2	282
名前_その他	66.5	75.5	70.4	211
神名	45.7	60.7	51.2	62
緯度経度	43.3	60.0	49.3	6
株指標	0.0	0.0	0.0	1
時間表現_その他	0.0	0.0	0.0	1

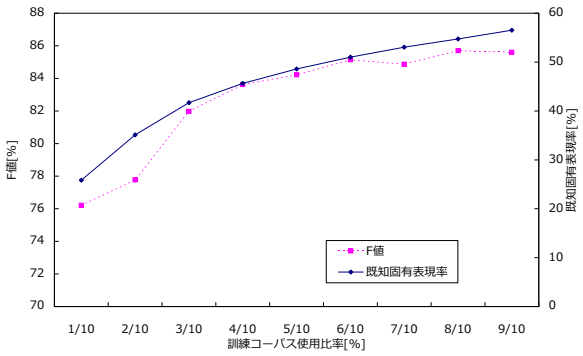

図 3: 新聞サブコーパス

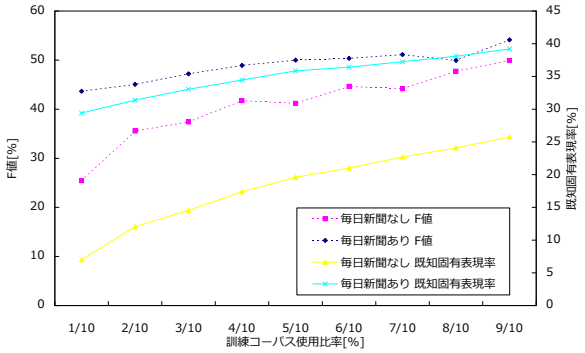

図 4: 白書サブコーパス

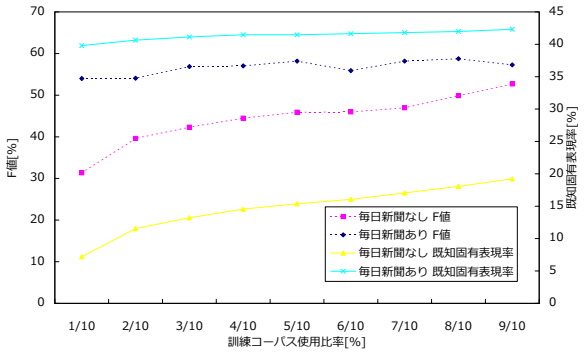

図 5: 書籍サブコーパス

練コーパスとして十分ではないと考えられる．そのため，各サブコーパスに新聞サブコーパスを加えたコーパスを用いて訓練した実験も行った．新聞サブコーパスを加えたことにより，サブコーパス単体で学習した場合に比べれば抽出性能は改善しているが，新聞サブコーパス同士の抽出性能よりは劣っている．これは，対象となるサブコーパスと新聞サブコーパスは文体などで異なっており，単純に新聞サブコーパスを追加しただけでは，その対象サブコーパスについて有効な学習が行えていないからだと考えられる．

5 おわりに

本研究では，拡張固有表現抽出を，形態素情報と文字種情報を素性とする文字列に対するチャンキング問題として定式化し，機械学習手法として CRF を用いた場合について，オンライン学習 (FOBOS) とバッチ学習 (L-BFGS) の学習に要する時間を比較した．その結果，全ての拡張固有表現タイプを十分に含むような大規模コーパスに対しては，オンライン学習 (FOBOS) が必要となるという見通しを得た．また，現状の拡張固有表現タグ付きコーパスは，統計的機械学習手法として CRF を用いた場合には，訓練コーパスとしての分量が十分とは言えないサブコーパスを含むことを示

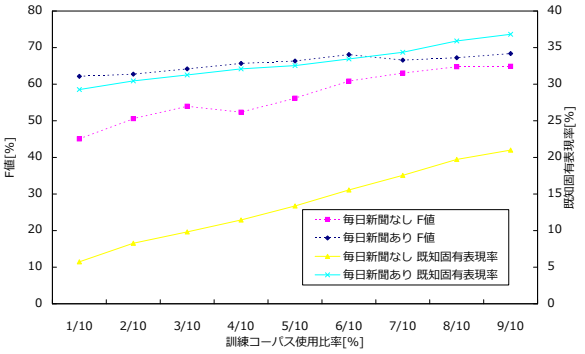

図 6: WEB サブコーパス

した．よって，今後は，拡張固有表現タグを含まないコーパスを併用する半教師あり学習などの手法の検討が必要である．

参考文献

[1] John Duchi and Yoram Singer. E cient learning using forward-backward spliting. In *NIPS2009*, 2009.

[2] Ralph Grishman and Beth Sundheim. Message understanding conference-6: a brief history. In *Proc. of the 16th COLING*, pp. 466–471, 1996.

[3] John La erty, Andrew McCallum, and Fernando Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of ICML*, pp. 282–289, 2001.

[4] D.C. Liu and J. Nocedal. On the limited memory method for large scale optimization. *Mathematical Programming B*, Vol. 45, No. 3, pp. 503–528, 1989.

[5] Satoshi Sekine and Yoshio Eriguchi. Japanese named entity extraction evaluation: analysis of results. In *Proc. of the 18th COLING*, pp. 1106–1110, 2000.

[6] Satoshi Sekine, Kiyoshi Sudo, and Chikashi Nobata. Extended named entity hierarchy. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, pp. 1818–1824, 2002.

[7] 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, Mar 2004.

[8] 新納浩幸, 関根聡. 拡張固有表現タガーの作成とその問題点の考察. 言語処理学会第 12 回年次大会発表論文集, pp. 105–108, 2006.

[9] 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, May 2004.

[10] 山崎誠, 前川喜久雄, 田中牧郎, 小椋秀樹, 柏野和佳子, 小磯花絵, 間淵洋子, 丸山岳彦, 山口昌也, 秋元祐哉, 稲益佐知子, 吉田谷幸宏. 代表性を有する現代日本語書き言葉コーパスの設計. 言語処理学会第 12 回年次大会発表論文集, pp. 440–443, 2006.

[11] 橋本泰一, 乾孝司, 村上浩司. 拡張固有表現タグ付きコーパスの構築. 情報処理学会研究報告, 第 2008–NL–188 巻, pp. 113–120. 社団法人情報処理学会, 2008.