ブートストラップによる低人手コスト日本語固有表現抽出

宇津呂 武仁
豊橋技術科学大学 情報工学系
utsuro@ics.tut.ac.jp

颯々野 学
富士通研究所
sassano@flab.fujitsu.co.jp

本論文では，人手によって作成された少量の初期固有表現リストと大量の人手未解析コーパスから，ブートストラップにより日本語固有表現抽出規則を学習する手法を提案し，その実験的評価結果を報告する．実験の結果，ブートストラップのサイクルを経るにしたがって，初期固有表現リストによる固有表現抽出の性能 (F 値) が徐々に改善されるという結果が得られた．この結果により，少量の初期知識と大量の人手未解析コーパスを用いたブートストラップ法のアプローチが，日本語の固有表現のまとめ上げの問題においても，ある程度有効に機能することが確認できた．

Low Manual Cost Japanese Named Entity Recognition by Bootstrapping

Takehito Utsuro
Department of Information and Computer Sciences,
Toyohashi University of Technology
utsuro@ics.tut.ac.jp

Manabu Sassano
Fujitsu Laboratories, Ltd.
sassano@flab.fujitsu.co.jp

Approaches to named entity recognition that rely on hand-crafted rules and/or supervised learning techniques have limitations in terms of their portability into new domains as well as in the robustness over time. For the purpose of overcoming those limitations, this paper evaluates named entity chunking and classification techniques in Japanese named entity recognition in the context of minimally supervised learning. This experimental evaluation demonstrates that the minimally supervised learning method proposed here improved the performance of the seed knowledge on named entity chunking and classification. We also investigated the correlation between performance of the minimally supervised learning and the sizes of the training resources such as the seed set as well as the unlabeled training data.

1 はじめに

固有表現抽出は，情報検索・抽出，機械翻訳，自然言語理解など自然言語処理の応用的局面における基礎技術として重要な技術の一つである．英語においては，特に米国において，MUC(Message Understanding Conference，例えば，MUC-7 [MUC98]) コンテストにおける課題の一つとして固有表現抽出がとりあげられ，集中的に研究が行なわれてきた．また，最近では，日本語においても，MET (Multilingual Entity Task, 例えば，MET-1 [Maiorano96] ，MET-2 [MUC98]) や IREX(Information Retrieval and Extraction Exercise) ワークショップ [IREX 実行委員会 99] などのコンテストにおいて，固有表現抽出が課題の一つに取り上げられている．

日本語解析においては，近年，高性能な形態素解析システムが利用可能となってきており，日本語固有表現抽出においても，形態素解析の結果得られる情報が非常に有用であるため，日本語固有表現抽出の研究においては，前処理として形態素解析を行なうことが標準的となっている．そして，形態素解析結果の形態素列に対して，人手で構築されたパターンマッチング規則や統計的学習によって得られた固有表現抽出規則を適用することにより，固有表現が抽出される [IREX 実行委員会 99]．人手で規則を記述するアプローチと，人手で構築された正解データから統計的手法により規則を学習するアプローチを比べると，後者のアプローチでは，共通の訓練用コーパスが公開されている場合に人手のコストをかけずに規則を自動学習できるという利点がある．また，対象とする分野や固有表現の定義が変わった場合でも，一度訓練用のコーパスを整備してしまえば，その訓練コーパスの傾向に沿った規則が自動学習できるという利点もある [Sekine98, Borthwick99, 内元 00, 颯々野 00b]．

一方で，固有表現の特質として，新聞記事など最も移り変わりの激しいテキストを抽出対象とする場合，分野や時代が変わると新たな固有表現が次々に出現し，固有表現抽出規則の更新が必要となるという深刻な問題が生じてくる．その場合，人手による規則のアプローチでは，人手で規則を更新する必要が生じ，また，人手で訓練用のコーパスを整備しそこから規則を自動学習するアプローチでも，その都度訓練用のコーパスを整備し直す必要が生じてくる．

このような問題に対する有効なアプローチとして，近年，人手によって作成された少数の規則もしくは少量の初期固有表現リストと大量の人手未解析コーパスから，Co-training[Blum98] などのブートストラップの考え方に基づいて固有表現抽出規則などを学習するというアプローチの研究が行なわれている [Collins99, Cucerzan99, Riloff99, 合原 00]．このアプローチの基本的な考え方は，現在位置の単語の綴りの情報と周囲の単語の綴りの情報など，性質の異なる二種類以上の情報に注目し，これらの情報が一つの固有表現において複数出現しているというコーパスの冗長性を利用して，性質の異なる二種類以上の情報を相互に補い合いながら，ブートストラップにより人手の労力を極力抑えて固有表現抽出規則を学習するというものである．

これらの先行研究をふまえて，本論文では，人手によって作成された少量の初期固有表現リストと大量の人手未解析コーパスから，ブートストラップにより日本語固有表現抽出規則を学習する手法を提案し，その実験的評価結果を報告する．一般に，人手解析済の訓練用コーパスを用いた教師あり学習により日本語固有表現抽出規則を学習する手法の先行研究 [Sekine98, Borthwick99, 内元 00, 颯々野 00b] によれば，日本語の固有表現抽出の問題の大部分は，形態素解析結果の形態素列に対して，一つもしくは複数の形態素のまとめ上げ処理を行ない，同時にまとめ上げられた形態素列がどの種類の固有表現を構成しているかを同定するプロセスにより有効に対処できることが分かっている．つまり，日本語の固有表現抽出問題は，本質的に形態素列の固有表現へのまとめ上げ問題である．これまで，ブートストラップの考え方に基づく先行研究 [Collins99, Cucerzan99, Riloff99, 合原 00] のいずれにおいても，まとめ上げ問題を扱ったものはなかった．したがって，本論文の焦点の一つは，少量の初期知識と大量の人手未解析コーパスを用いたブートストラップ法のアプローチが，まとめ上げ問題にどの程度有効であるかを検証する点にあると言える．

以下，本論文で用いたブートストラップ法を簡単に説明する．まず，人手により少量の初期固有表現リストを作成し，この初期固有表現リストを固有表現抽出規則の形式に変換する．次に，この固有表現抽出規則を大量の人手未解析コーパスに適用し，その適用結果を訓練データとみなして，通常の教師あり学習アルゴリズムにより新たな固有表現抽出規則を学習する．以下，同様に，大量の人手未解析コーパスに対して，規則学習と規則適用を繰り返す．ここで，ブートストラップ方式における教師あり学習アルゴリズムとしては，従来，決定リスト学習もしくはそれと類似の手法が用いられてきた [Yarowsky95, Collins99, Cucerzan99] ため，本研究においても，決定リスト学習を用いる．

このブートストラップ法によって，実際に日本語固有表現抽出規則の学習の実験を行なったところ，ブートストラップのサイクルを経るにしたがって，初期固有表現リストによる固有表現抽出の性能 (F 値) が徐々に改善されるという結果が得られた．この結果により，少量の初期知識と大量の人手未解析コーパスを用いたブートストラップ法のアプローチが，日本語の固有表現のまとめ上

表 1: 固有表現まとめ上げ状態の表現法

固有表現タグ		<ORG>			<LOC>			<LOC>		
形態素列	$\cdots$ M	M	M		M	M	M	M	M	$\cdots$
Inside/Outside 法		O	ORG_I	O	LOC_I	LOC_I	LOC_I	LOC_B	O	
Start/End 法		O	ORG_U	O	LOC_S	LOC_C	LOC_E	LOC_U	O	

げの問題においても，ある程度有効に機能することが確認できた．

2 固有表現まとめ上げおよび分類の教師あり学習

本節では，本論文で用いた日本語固有表現まとめ上げおよび分類の教師あり学習手法について簡単に述べる．

2.1 固有表現まとめ上げ状態の表現法

固有表現まとめ上げの際のまとめ上げ状態の表現法としては，Inside/Outside 法および Start/End 法を用いた [Sassano00a, 颯々野 00b]．これらの二種類の方法により日本語固有表現のまとめ上げを行なう様子を表 1 に示しておく．

2.1.1 Inside/Outside 法

この方法は英語の単名詞句まとめ上げでよく用いられる方法の一つである [Ramshaw95]．単名詞句まとめ上げの場合には，まとめ上げ状態として以下の三種類の状態を設定する．

O – 現在位置の単語はどの単名詞句にも含まれない．

I – 現在位置の単語は一つの単名詞句の一部である．

B – 現在位置の単語は，ある単名詞句の直後の位置する別の単名詞句の先頭の単語である．

本論文では，この方法を固有表現まとめ上げおよび分類に適用し，状態 I および B をそれぞれ八種類の固有表現タイプに細分類する．結果として，この表現法では，固有表現まとめ上げ状態として，$2 \times 8 + 1 = 17$ の状態を設定する．

2.1.2 Start/End 法

この方法は，日本語固有表現抽出の既存の手法 [Sekine98, Borthwick99, 内元 00] において用いられた方法で，各固有表現タイプについて，以下の四種類のまとめ上げ状態を設定する．

S – 現在位置の形態素は，一つ以上の形態素から構成される固有表現の先頭の形態素である．

C – 現在位置の形態素は，一つ以上の形態素から構成される固有表現の先頭・末尾以外の中間の形態素である．

E – 現在位置の形態素は，一つ以上の形態素から構成される固有表現の末尾の形態素である．

U – 現在位置の形態素は単独で一つの固有表現を構成する．

また，固有表現を構成しない形態素のための状態として以下の状態を設定する．

O – 現在位置の形態素はどの固有表現にも含まれない．

結果として，この表現法では，固有表現まとめ上げ状態として，$4 \times 8 + 1 = 33$ の状態を設定する．

2.2 周囲の形態素のモデル化

本論文では，現在位置の形態素に対して固有表現のまとめ上げ状態を付与する際に，周囲のどれだけの形態素を考慮するか，つまり周囲の形態素をどのようにモデル化するかについては，簡単のため，現在位置の形態素の左側および右側について常に一つずつの形態素のみを考慮する 3 グラムモデル [Sassano00a, 颯々野 00b] を用いる．

2.3 決定リスト学習

教師あり学習の方法としては，[Yarowsky94] の決定リスト学習を用いる．決定リストは，ある素性 F のもとでクラス D を決定するという規則を優先度の高い順にリスト形式で並べたもので，適用時には優先度の高い規則から順に適用を試みていく．[Yarowsky94] の決定リスト学習の方法においては，クラス D の正解付データから，素性 F が存在する $(F = 1)$ という条件のもとでクラス D が $D{=}x$ となる条件付確率 $P(D{=}x \mid F{=}1)$ を計算し，この条件付確率を用いて以下の手順で決定リストを構成する．

1. ある素性 F が存在する $(F{=}1)$ という条件のもとでの条件付確率 $P(D{=}x \mid F{=}1)$ の値の大きさが一位のクラス x_1 と二位のクラス x_2 の間で，以下の対数尤度比を計算しこの値を確信度とする．

$$\log_2 \frac{P(D{=}x_1 \mid F{=}1)}{P(D{=}x_2 \mid F{=}1)}$$

その結果，確信度が大きい順に素性 F とクラス D の組を並べる．ただし，このときの確信度は，クラス $D{=}x$ の確率 $P(D{=}x)$ の値の大きさが一位のクラス x_1 と二位のクラス x_2 の間で以下の対数尤度比を計算して得られる値

$$\log_2 \frac{P(D{=}x_1)}{P(D{=}x_2)}$$

を下限値とする．

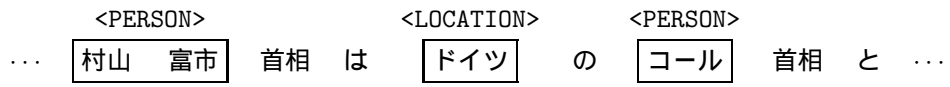


図 1: ブートストラップの基本的考え方: コーパス中の冗長性の利用の例

2. 決定リストの最終行は “default” を表し，クラス $D=x$ の確率 $P(D=x)$ の値の大きさが一位のクラス x_1 を与える．

2.4 素性およびクラス

各形態素の素性として用いるのは，語彙と品詞の組，文字種と品詞の組，品詞の三種類である．品詞は，形態素解析システム BREAKFAST[颯々野 97] の約 300 種類を用いる．また，文字種は，平仮名・片仮名・漢字・数字・英語アルファベット・記号，およびそれらの組合わせを用いる．決定リスト学習の際には，参照する全ての形態素 (本論文では，現在位置の形態素およびその左側および右側について一つずつの形態素) の素性の組合わせを素性候補とする．その際，現在位置の形態素については何らかの素性を考慮する必要があるが，その他の周囲の文脈の形態素の素性については，情報を省略することが可能である．クラスは，Inside/Outside 法あるいは Start/End 法に応じて，2.1 節で述べた各固有表現まとめ上げ状態がその値となる．

3 ブートストラップによる低人手コスト固有表現抽出

3.1 基本的考え方: 人手未解析コーパス中の冗長性の利用

ブートストラップの基本的考え方は，人手未解析コーパス中において，一つの固有表現の周囲に複数の手がかりが存在するという冗長性を利用することである．以下では，例を用いてこの考え方を説明する．

まず，図 1 に示すような形態素列が人手未解析コーパス中に含まれているとする．図 1 には，「村山 富市」(<PERSON>)「ドイツ」(<LOCATION>)「コール」(<PERSON>) といった，正しい固有表現の情報も記されているが，これらは人手未解析コーパス中には存在していない．

ここで，初期固有表現の情報として「村山 富市」(<PERSON>) のみが与えられたとする．すると，この初期固有表現の情報と図 1 の人手未解析コーパス中の出現パターンから，以下のような固有表現抽出規則が学習される．

規則: if 「X 首相」 then X is 固有表現 & <PERSON>

次に，この規則を，同じ人手未解析コーパスに適用していくことにより，「コール」(<PERSON>) という固有表現を抽出することができる．つまり，初期固有表現と人手未解析コーパスを用いることにより，初期固有表現には含まれていなかった「コール」(<PERSON>) という固有表現を抽出できたことになる．この場合は，「村山 富市」(<PERSON>) という固有表現の位置において，初期固有表現の情報と後続の文脈の「首相」という手がかりの二つの情報が冗長に存在していたことによって，ブートストラップが可能となった．

3.2 アルゴリズム

ブートストラップのアルゴリズムを以下で説明する．このアルゴリズムは，人手解析済コーパスを用いず，初期固有表現および人手未解析コーパスのみを用いて固有表現抽出規則を学習する．

1. 初期処理

初期固有表現リストの作成

まず，形態素および形態素列の頻度リストと，固有表現の定義を参照しながら，人手で高頻度固有表現のリストを作成し，これを初期固有表現リストとする[1] ．

初期固有表現リストの決定リストへの変換

次に，初期固有表現リストの各固有表現を，それぞれ，決定リストの形式の固有表現抽出規則に変換する．複数の形態素から構成される固有表現の場合は，複数の規則に変換される．いずれの規則も，周囲の形態素は考慮せずに，現在位置の形態素の語彙情報 (および可能であれば品詞情報) を条件として，固有表現まとめ上げ状態を出力するという形式で記述される．また，各規則は初期決定リスト中では同一の確信度を与えられる．

2. ブートストラップ

ブートストラップのプロセスは，以下の決定リスト適用および学習のサイクルを繰り返すことにより行なわれる．また，各サイクルにおいて，別途用意した人手解析済評価用コーパスに対する適合率・再現率・F 値の変化を測定する．また，訓練用人手未解析コーパスに正解の固有表現情報が付与されている場合は，訓練用人手未解析コーパスに対する適合率・再現率・F 値の変化も測定する．

人手未解析コーパスへの決定リストの適用

得られている決定リストを人手未解析コーパスに適用し固有表現を同定する．この際，決定リストに対する確信度の閾値は設けず，全ての規則を用いる[2] ．

[1] このとき，高頻度固有表現でなく，頻度に関して無作為に固有表現を収集し固有表現リストを作成した場合の性能についても，4.2.2 節でその結果を説明する．

[2] これは経験的に得られた設定である．様々な設定で実験を行なった結果では，確信度の閾値を設けた場合は，ブートストラップのサイクルを経ても F 値の上昇率が悪くなった．一般に教師あり学習の場合

表 2: 教師あり学習の実験結果

		n 形態素 対 1 固有表現				
		$n \geq 1$	$n = 1$	$n = 2$	$n = 3$	$n \geq 4$
Inside/Outside 法	F 値 ($\beta = 1$)	**72.9**	75.9	**79.7**	**51.4**	29.2
	(適合率)	(72.6)	(68.7)	(78.4)	(74.7)	(87.5)
	(再現率)	(73.1)	(84.7)	(81.1)	(39.2)	(17.5)
Start/End 法	F 値 ($\beta = 1$)	72.7	**76.6**	79.6	43.7	**29.6**
	(適合率)	(76.7)	(72.7)	(82.4)	(79.6)	(95.5)
	(再現率)	(69.0)	(81.0)	(77.0)	(30.1)	(17.5)

適用結果からの決定リスト学習

人手未解析コーパスに決定リストを適用した結果を，通常の教師あり学習に用いる訓練データとみなして，固有表現抽出のための決定リストを学習する．これは，誤りを含む訓練データからの決定リストの学習に相当する．

4 実験および評価

本節では，IREX ワークショップの固有表現抽出タスクにおいて配布された CRL(郵政省通信総合研究所) 固有表現データ (毎日新聞 1,174 記事の固有表現をタグ付け)，および本試験データのうちの一般ドメインのもの (毎日新聞 71 記事の固有表現をタグ付け) を用いた行なった実験結果について述べる．

4.1 固有表現まとめ上げ状態表現法の比較: 予備実験結果

まず，固有表現まとめ上げ状態の表現法として用いる Inside/Outside 法と Start/End 法の二つを比較するために，予備実験として教師あり学習およびブートストラップ法の性能の比較を行なった．

4.1.1 教師あり学習

IREX ワークショップの CRL 固有表現データを訓練データとし本試験データを評価データとして，教師あり学習の評価を行なった．ここで述べる実験結果は，教師あり学習に関する別の詳細な報告 [颯々野 00b] からの抜粋である．表 2 に，Inside/Outside 法と Start/End 法のそれぞれについて，固有表現を構成する形態素の数ごとに測定した F 値 ($\beta = 1$)・適合率・再現率の値を示す．また，Inside/Outside 法および Start/End 法の F 値のうち，より高い方の値を太字で示す．

全体の性能では，Inside/Outside 法の方が若干高い F 値を示しており，また，三形態素から構成される固有表現の場合は，Inside/Outside 法の方が 8 ポイントほど高い F 値を示している．一方，Start/End 法の方は，一形態素および四形態素以上から構成される固有表現の場合に若干高い F 値を示している．また，Inside/Outside 法が再現率で上回っているのに対し，Start/End 法は適合率で上回っているが，これは，固有表現まとめ上げ状態の数が少ない Inside/Outside 法がより一般化することにより再現率を重視する傾向にあり，逆に固有表現まとめ上げ状態の数が多い Start/End 法がより正確にモデル化し適合率を重視する傾向にあることを示している．

4.1.2 ブートストラップ法

次に，ブートストラップ法のシミュレーションとして，教師あり学習で得られた固有表現抽出規則の上位のものを初期決定リストして，ブートストラップ法により固有表現抽出規則の学習を行ない，Inside/Outside 法と Start/End 法の比較を行なった．

まず，IREX ワークショップの CRL 固有表現データを半分ずつに分割し，その一方を訓練データとして，教師あり学習により固有表現抽出用の決定リストを学習した．そして，得られた決定リストの上位 100 および 500 規則を初期決定リストとして[3] ，残りの半分の CRL 固有表現データを人手未解析のブートストラップ用訓練データとみなして，ブートストラップのサイクルを実行した．Inside/Outside 法と Start/End 法の双方について，このブートストラップのサイクルの過程において，IREX ワークショップの本試験データに対して測定した F 値 ($\beta = 1$) の変化の様子を図 2 に示す．

図 2 の結果から，初期決定リストの規則数が同じ場合には，まず，初期段階の F 値において Inside/Outside 法が Start/End 法を上回っていることがわかる．また，この傾向は，ブートストラップのサイクルを経た後も変わらず，サイクルを通しての最高の F 値についても Inside/Outside 法が Start/End 法を上回っている．この傾向は，Start/End 法の方が固有表現まとめ上げ状態の数が多いため，Inside/Outside 法と同等の性能を示すためにはより多くの規則が必要となることが原因である．

は，決定リストの確信度に適当な閾値を設けた方がよい性能が得られる [Sassano00a] が，本研究の設定におけるブートストラップにおいては逆の結果となった．

[3] Inside/Outside 法と Start/End 法のいずれにおいても，教師あり学習により得られた決定リストの規則数は約 60,000 個で，このうち上位の 500 個はいずれも一意にまとめ上げ状態を決定する規則 (すなわち，平滑化を施さない場合の，素性を条件とするクラスの条件付確率 $P(D{=}x \mid F{=}1)$ の値が 1 に等しい) であった．

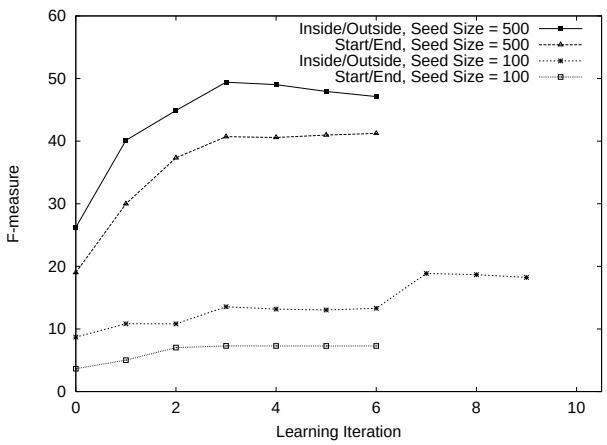


図 2: ブートストラップ法における固有表現まとめ上げ状態表現法の比較: 予備実験結果

4.1.3 考察

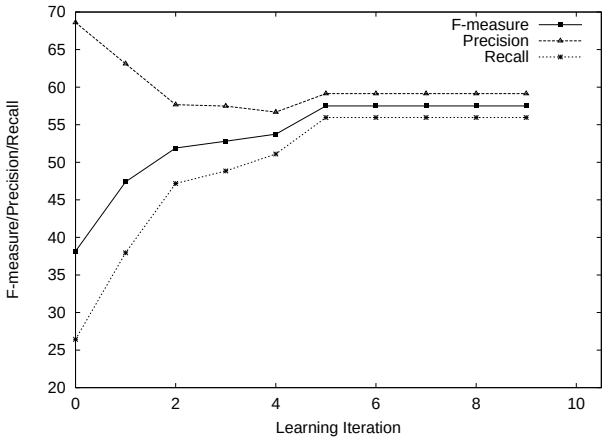
前節でも述べたが，Inside/Outside 法と Start/End 法は，固有表現まとめ上げ状態の数が異なることから，固有表現まとめ上げにおける汎化能力が大きく異なっている．Inside/Outside 法では，いずれの場合でも固有表現の内部であることを表すまとめ上げ状態 “I” が用いられる．したがって，教師あり学習やブートストラップにおいても，長さの異なる固有表現から学習された規則がお互いに区別なく適用されることになり，結果的に，モデルが固有表現の長さの違いを吸収していることになる．一方，Start/End 法では，一つの形態素から構成される固有表現と二つ以上の形態素から構成される固有表現は，全く異なるまとめ上げ状態を経て抽出される．したがって，教師あり学習やブートストラップの場合にも，長さの異なる固有表現から学習された規則は必ず区別されて適用されるため，モデルが固有表現の長さの違いを吸収することはできていない．

この汎化能力の違いは，人手未解析コーパスを用いるブートストラップ法の場合のようにデータの過疎性の問題が起き易い場合には，より重要な要因となると考えられる．ブートストラップ法における Inside/Outside 法と Start/End 法の長短を厳密に見極めるには，徹底した調査・分析が必要であるが，本稿の範囲においては，より汎化能力の高い Inside/Outside 法を用いることとし，次節では Inside/Outside 法を用いた場合のブートストラップ法の実験結果について詳細に述べる．

4.2 Inside/Outside 法を用いた場合のブートストラップ法の実験結果

固有表現まとめ上げ状態の表現法として Inside/Outside 法を用いた場合の，ブートストラップ法の実験を行なった．ここでは，便宜上，初期固有表現リストに含まれる固有表現としては，一つの形態素から構成される固有表現のみを用いた．まず，4.1.2 節の場合と同様に，IREX ワークショップの CRL 固有表現データを半分ずつに分割し，その一方から高頻度初期固有表現を抜き出して初期固有表現リストを作成した．そして，残りの半分の CRL 固有表現データ (の全てまたはその一部) を人手未解析のブートストラップ用訓練データとみなして，ブートストラップのサイクルを実行した．また，各サイクルにおいて，ブートストラップ用訓練データおよび IREX ワークショップ本試験データに対して，ブートストラップ法により学習された固有表現抽出規則の性能を測定した．

(a) 固有表現全体での適合率・再現率・F 値



(b) 固有表現の形態素長ごとの F 値

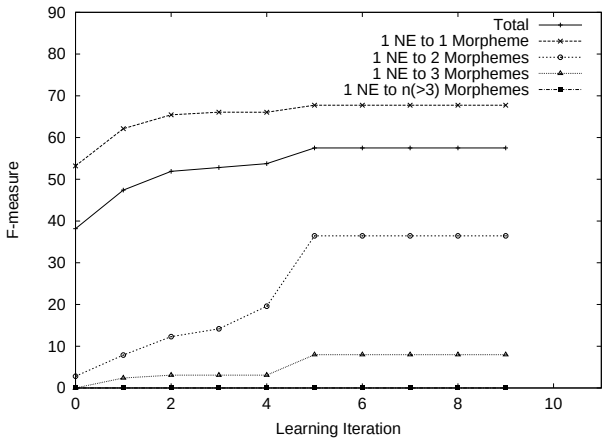
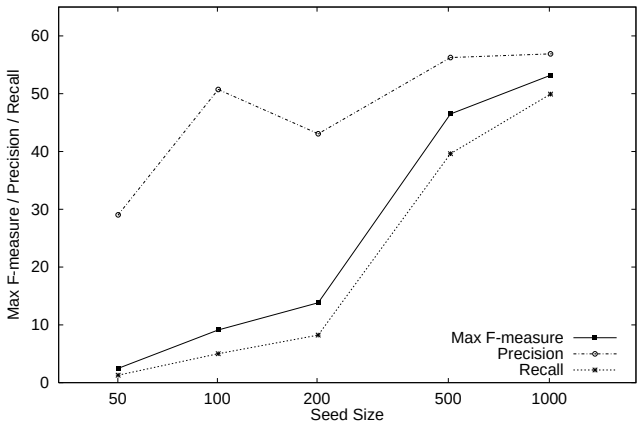


図 3: ブートストラップ法: 訓練時の性能変化

4.2.1 訓練時の適合率・再現率・F 値の変化

初期固有表現として頻度上位 200 個を用い，ブートストラップ用訓練データとして残りの半分の CRL 固有表現データ全てを用いた場合の適合率・再現率・F 値の変化を図 3 (a) に，また，固有表現の形態素長ごとの F 値の変化を図 3 (b) に示す．ここでは，ブートストラップ用訓練データに対する性能の変化を示しているが，IREX ワークショップの本試験データに対する性能の変化もほぼ同様の傾向を示す．まず，図 3 (a) の適合率・再現率・F 値の変化は，適合率が徐々に低下する一方で再現率が急激に上昇し，結果として F 値も上昇するという，ブートスト

(a) 頻度に関して無作為に初期固有表現を選んだ場合



(b) 高頻度順に初期固有表現を選んだ場合

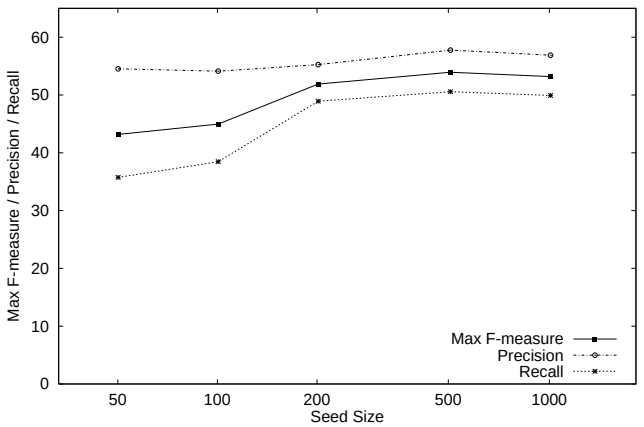
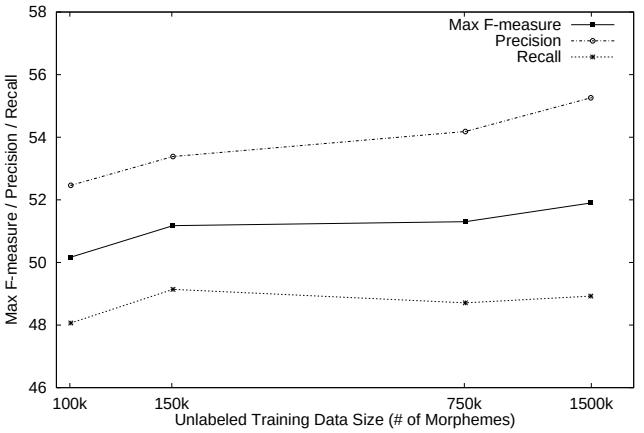


図 4: 初期固有表現数と性能の相関

(a) 評価データに対する性能



(b) 訓練データに対する性能

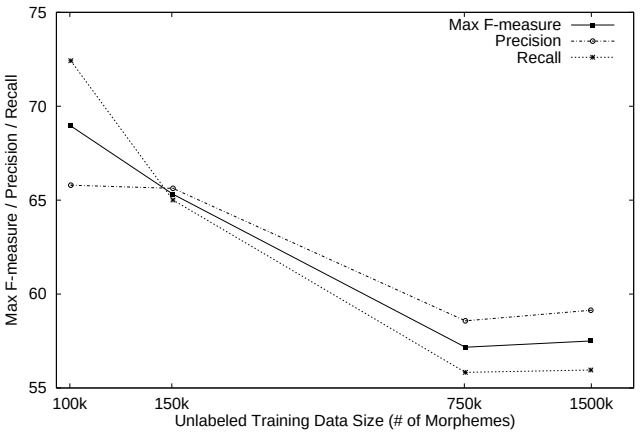


図 5: 訓練データサイズと性能の相関

ラップ法の典型的な変化の様子を示している．また，図 3 (b) の結果で興味深いのは，初期固有表現リストは一つの形態素から構成される固有表現だけを集めていたにも関わらず，ブートストラップのサイクルを通じて二形態素および三形態素から構成される固有表現の F 値が上昇している点である．これは，4.1.3 節で述べた Inside/Outside 法の汎化能力の高さを示す結果であると言える．

4.2.2 初期固有表現数および訓練データサイズと性能の相関

次に，初期固有表現数および訓練データの人手未解析コーパスのサイズを変化させ，ブートストラップ法によって得られる性能との相関を調べた結果を述べる．

まず，初期固有表現数を 50，100，200，500，1000 と変化させてブートストラップのサイクルを実行し，各初期固有表現数の設定において，IREX ワークショップの本試験データに対して測定した最高の F 値 (およびそのときの適合率・再現率) をプロットした結果を図 4 に示す．そのうち，図 4 (a) は，頻度に関して無作為に初期固有表現を選んだ場合の結果であり，図 4 (b) は，高頻度順に初期固有表現を選んだ場合の結果である．図 4 の横軸はいずれも対数スケールで書かれており，図 4 (a) においては，最高の F 値の値が初期固有表現数に対してほぼ対数線形に増加している．この傾向は，ブートストラップを用いた多言語固有表現抽出の先行研究 [Cucerzan99] における実験結果の傾向と一致している．また，図 4 (a) と (b) のうち，初期固有表現数が少ない部分の最高の F 値を見比べると，図 4 (b) の方が圧倒的に高い性能を示していることが分かる．これはつまり，同じ数の初期固有表現を用いるのであれば高頻度なものを用いた方がはるかに効果的であることを示している．また，図 4 (b) においては，初期固有表現数が 500 の点において最高の F 値の増加がほぼ飽和している．これは，低頻度な固有表現を初期固有表現リストに追加してリストのサイズを大きくしても，性能が悪化することを意味している．

次に，訓練データの人手未解析コーパスのサイズを変化させ，頻度上位 200 個を初期固有表現としてブートストラップのサイクルを実行し，各訓練データサイズの設定における最高の F 値 (およびそのときの適合率・再現率) をプロットした結果を図 5 に示す．そのうち，図 5 (a) は，IREX ワークショップの本試験データに対する性能をプロットしたもので，一方，図 5 (b) は，ブートストラップのサイクルに用いている訓練データに対する性能をプ

ロットしたものである．ここで，図 5 の横軸は，訓練データの形態素数の対数スケールで書かれている．図 5 (a) では，訓練データサイズの増加に伴って，最高の F 値における適合率が上昇しており，これが原因で最高の F 値もわずかずつ上昇している．この傾向は，訓練データの人手未解析コーパスのサイズが大きければ大きいほどより正確な規則が学習できるという先行研究 [Cucerzan99] の主張と一致している．一方，図 5 (b) では，訓練データサイズの増加に伴って，ブートストラップのサイクルに用いている訓練データに対する性能は悪化している．これは，訓練データサイズが増加するのに伴って，低頻度で抽出の困難な固有表現がより多く含まれるようになったためと考えられる．

5 おわりに

本論文では，人手によって作成された少量の初期固有表現リストと大量の人手未解析コーパスから，ブートストラップにより日本語固有表現抽出規則を学習する手法を提案し，その実験的評価結果を報告した．実験の結果，ブートストラップのサイクルを経るにしたがって，初期固有表現リストによる固有表現抽出の性能 (F 値) が徐々に改善されるという結果が得られた．この結果により，少量の初期知識と大量の人手未解析コーパスを用いたブートストラップ法のアプローチが，日本語の固有表現のまとめ上げの問題においても，ある程度有効に機能することが確認できた．今後は，実験の設定のうちまだ行なえていないもの (固有表現まとめ上げ状態の表現法として Start/End 法を用いた場合，および，初期固有表現として任意の形態素長の固有表現を用いた場合) についても実験を行ない，その性能を分析する．また，人手解析済コーパスを訓練データとする教師あり学習においては，学習アルゴリズムとして最大エントロピー法を用い，学習時に，現在位置の形態素が，いくつの形態素から構成される固有表現の一部であるかを考慮して学習を行なう方法 (可変長モデル) が最も高性能であることがわかっている [颯々野 00b] ので，ブートストラップにおいてもその方式の効果を検証する．また，実用的観点からは，ブートストラップ方式の利点を生かしながら，人手の介入を最小限に抑え，かつ，人手解析済コーパスを訓練データとする教師あり学習を用いた場合と同程度の性能を持った方式を開発する必要があるであろう．

謝辞

本研究の成果の多くは，筆者らが米国ジョンズホプキンス大学計算機科学科に客員研究員として滞在中に得られたものである．本研究に対し多くの貴重なコメントを頂いた同大学 David Yarowsky 教授に感謝する．

参考文献

[Blum98] Blum, A. and Mitchell, T.: Combining Labeled and Unlabeled Data with Co-Training, *Proceedings of the 11th Annual Conference on Computational Learning Theory*, pp. 92–100 (1998).

[Borthwick99] Borthwick, A.: A Japanese Named Entity Recognizer Constructed by a Non-Speaker of Japanese, IREX ワークショップ予稿集, pp. 187–193 (1999).

[Collins99] Collins, M. and Singer, Y.: Unsupervised Models of Named Entity Classification, *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pp. 100–110 (1999).

[Cucerzan99] Cucerzan, S. and Yarowsky, D.: Language Independent Named Entity Recognition Combining Morphological and Contextual Evidence, *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pp. 90–99 (1999).

[合原 00] 合原博, 宮田高志, 松本裕治：医学生物学分野からの専門用語の抽出・分類, 情報処理学会研究報告, Vol. 2000, No. (2000–NL–135), pp. 41–48 (2000).

[IREX 実行委員会 99] IREX 実行委員会 (編): IREX ワークショップ予稿集 (1999).

[Maiorano96] Maiorano, S.: The Multilingual Entity Task (MET): Japanese Results, *Proceedings of TIPSTER PROGRAM PHASE II*, pp. 449–451 (1996).

[MUC98] MUC: *Proceedings of the 7th Message Understanding Conference (MUC-7)* (1998).

[Ramshaw95] Ramshaw, L. and Marcus, M.: Text Chunking using Transformation-Based Learning, *Proceedings of the 3rd Workshop on Very Large Corpora*, pp. 83–94 (1995).

[Riloff99] Riloff, E. and Jones, R.: Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping, *Proceedings of the 16th AAAI*, pp. 474–479 (1999).

[颯々野 97] 颯々野学, 斎藤由香梨, 松井くにお：アプリケーションのための日本五形態素解析システム, 言語処理学会第 3 回年次大会論文集, pp. 441–444, 言語処理学会 (1997).

[Sassano00a] Sassano, M. and Utsuro, T.: Named Entity Chunking Techniques in Supervised Learning for Japanese Named Entity Recognition, *Proceedings of the 18th COLING*, pp. 705–711 (2000).

[颯々野 00b] 颯々野学, 宇津呂武仁：統計的日本語固有表現抽出における固有表現まとめ上げ手法とその評価, 情報処理学会研究報告, Vol. 2000, No. (2000–NL–139) (2000).

[Sekine98] Sekine, S., Grishman, R. and Shinnou, H.: A Decision Tree Method for Finding and Classifying Names in Japanese Texts, *Proceedings of the 6th Workshop on Very Large Corpora*, pp. 148–152 (1998).

[内元 00] 内元清貴, 馬青, 村田真樹, 小作浩美, 内山将夫, 井佐原均：最大エントロピーモデルと書き換え規則に基づく 固有表現抽出, 自然言語処理, Vol. 7, No. 2, pp. 63–90 (2000).

[Yarowsky94] Yarowsky, D.: Decision Lists for Lexical Ambiguity Resolution: Application to Accent Restoration in Spanish and French, *Proceedings of the 32nd Annual Meeting of ACL*, pp. 88–95 (1994).

[Yarowsky95] Yarowsky, D.: Unsupervised Word Sense Disambiguation Rivaling Supervised Methods, *Proceedings of the 33rd Annual Meeting of ACL*, pp. 189–196 (1995).