E-034

決定木の逐次学習による固有表現抽出

Named-Entity Recognition based on Iterative Decision-Tree Learning

工藤 嘉晃† **相薗 敏子**†
Yoshimitsu Kudoh Toshiko Aizono

1. はじめに

固有表現（Named-Entity, 以下 NE）抽出は，情報検索・抽出，機械翻訳，質問応答システムにおいて，重要な基盤技術の一つである[1]．NE の概念は，1990 年前後に米国で情報抽出の研究を主導していたＭＵＣ（Message Understanding Conference）[2]において生まれた．NE の種類としては，人名，地名，組織名，および人工物名（製品名や法律名など）の固有名詞的表現，時間表現および数値表現がある[3].

従来方式において NE は，人手で構築されたパターンマッチング規則や，機械学習や統計的学習によって生成された NE 抽出規則（以下，抽出規則）を適用することにより，文書データから抽出されている．前者の規則を正確に人手で記述するためには，対象文書に対する領域知識と十分な作業時間を必要とする．これに対して後者の抽出規則は，教師データとなる大量の文書データが事前に存在する場合には，低いコストで自動的に生成できるという利点がある．近年，NE 抽出コンテストを目的として大量の教師データが公開されているため[4]，抽出規則を自動的に生成する研究が盛んに行われている[5][6]．しかしながら，実問題において事前に教師データが用意されていることは少ない．そのため，膨大な作業コストをかけて，教師データを構築しなければならない．

この問題に対処するために，ブートストラップ法の考え方[9]に基づいて人手で作成した少数の抽出規則もしくは少量の NE を用いて，大量の文書データから抽出規則を逐次的に学習する方式が提案されている[7][8]．宇津呂らの方式[7]では，決定リスト[10]を抽出規則として用いている．まず NE 抽出を行う前に，作業者が正例として少量の NE とそれらに対応した初期の抽出規則を作成する．次に，抽出規則を文書データに適用して NE 候補を抽出し，それらを新たに正例として抽出規則を学習する．このような抽出規則の学習を繰り返し行うことで，その規則の精度を向上させる．最後に学習した抽出規則を用いて，文書データから NE を抽出する．この方式は，十分な教師データがなくても高精度の抽出規則を学習し，効率的に NE を抽出することができる．

ブートストラップ法を採用した従来研究をふまえて本稿では，実用的な観点から二つの機能について検討する．一つは，正例と負例から抽出規則を学習する機能である．一般に正例のみからの学習では，規則が過剰に汎化され，抽出誤りが起こりやすくなる．これによる精度の低下を防ぐために，正例と共に負例を用いた学習を逐次的に実行する．もう一つの機能は，NE 候補のチェックにおいて作業者を支援するための機能である．

抽出規則の学習に正例と負例を用いる場合，抽出された NE 候補を正例とすると，NE 候補以外の単語はすべて負例となる．つまり，文書データ中には膨大な数の負例が存在することになる．一般に，正例に比べ負例が極端に多い場合は，精度の良い抽出規則を得ることは困難となる[11]．そこで，作業者が適切な負例を選択する必要がある。適切な負例の一つとして，抽出規則が NE 候補として誤って抽出した NE 以外の単語が挙げられる．これらの単語は，抽出規則の誤りを正して精度の向上に寄与する有効な負例となる．そのため作業者は，NE 候補に対して NE かどうか判別しなければならない．この判別における作業者の負担を軽減するために，判別の基準となる確信度という評価値を作業者に提示する．

本稿では，正例と負例から学習される決定木[11]を抽出規則として適用し，さらに抽出された NE 候補に対する作業者の判別コストを抑えるために確信度を利用した NE 抽出方式を提案する．また，本方式を用いた評価実験について報告する．

2. 提案方式

2.1 C4.5 による NE 抽出規則の学習

本方式では，決定木を学習するために C4.5[11]を用いる．そのためには，文書データを C4.5 に入力可能なデータ形式に変換する必要がある．本方式では，データの属性として NE 抽出で一般的に用いられる前後 n 単語の文字列，品詞，文字種を用いる（ただし，n は正の整数であり，その値は実験的に決定する）．

一例として，会社名“日立”や製品名“ＪＰ１”が正例として入力された場合，「日立の運用管理ツールＪＰ１を提案する．」という文書には，「<ORG>日立</ORG>の運用管理ツール<ARTIFACT>JP1</ARTIFACT>を提案する．」というように，組織名（<ORG>）や人工物名（<ARTIFACT>）を表すタグが付加される．そして，この文書に対して形態素解析を行う．形態素解析後の単語列「<ORG>日立</ORG>／運用／管理／ツール／<ARTIFACT>JP1</ARTIFACT>／提案」において，正例である“日立”や“JP1”を図 2 (a)のようなデータに変換する．組織名の正例である“日立”では，その後ろの単語“運用”，その品詞“名詞”およびその文字種“漢字”がデータの属性の値となる．

2.2 NE 抽出規則の逐次学習

ブートストラップ法に基づく従来の NE 抽出方式において正例と負例から抽出規則を学習する場合，NE 候補の判別が問題となる．これに対し本稿では，ブートストラップ法に NE 候補判別ステップを含めた NE 抽出方式を検討した．その結果，少量の教師データを繰り返し学習アルゴリ

† (株)日立製作所 中央研究所

ズムに与えて 抽出規則を逐次的に学習する方式を提案する．本方式を実装したシステムの処理の流れは次のようになる（図 1）．

(1) 作業者が既知の NE を正例として抽出規則学習部（以下，学習部）に入力する．
(2) 学習部が，文書データ中に出現する単語を負例候補とし，頻度順に作業者に提示する（2.3.1節参照）．
(3) 作業者が負例候補の中から NE ではないものを負例として選択し，学習部に入力する．
(4) 学習部は正例または負例を含む文書データを C4.5 に入力可能なデータ形式に変換し，教師データとする．教師データから決定木を学習し，それらを抽出規則に変換して規則の確信度（2.3.2節参照）とともに作業者に提示する．一例として図 2 (a)の教師データが C4.5 に入力されると，図 2(b)に示すような決定木を生成が生成される．
(5) 作業者が確信度の低い抽出規則を削除し，残った規則をデータベースに登録する．
(6) 抽出規則により NE 候補を抽出し，NE 候補の確信度（2.3.3節参照）とともに作業者に提示する．
(7) システムが提示した NE 候補に対して，作業者が最終的な判別を行い，NE を NE 辞書に登録する．

作業者は，NE 辞書に登録された NE を新たに正例としてシステムに入力し，ステップ 1 からステップ 7 を実行することで，新たに NE を抽出することが可能となる．

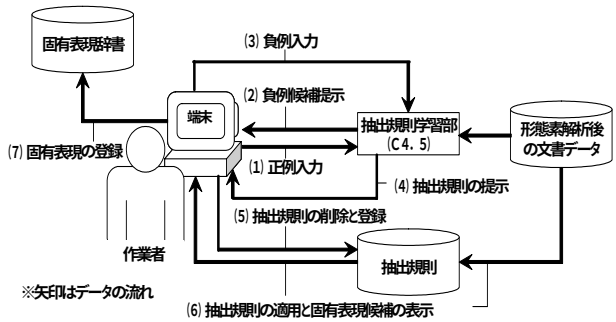

図 1　NE 抽出規則の逐次学習

	…	前1単語			後1単語			…
クラス	…	文字列	品詞	文字種	文字列	品詞	文字種	…
“正例(組織名)”	…	—	—	—	“運用”	“名詞”	“漢字”	…
“正例(製品名)”	…	“ツール”	“名詞”	“カタカナ”	“提案”	“動詞”	“漢字”	…
“負例”	…	“日立”	“名詞”	“漢字”	“管理”	“名詞”	“漢字”	…
“負例”	…	“管理”	“名詞”	“漢字”	“JP1”	“不明”	“英数字”	…
…	…	…	…	…	…	…	…	…

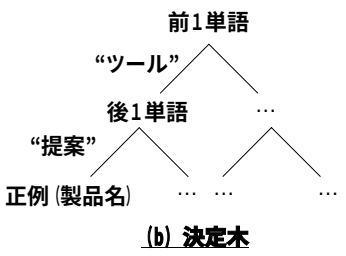

図 2　教師データと決定木の例

2.3　学習部の処理

2.3.1　負例の決定

本方式では，正例として与えた NE 以外の単語はすべて負例候補とするので，文書データには膨大な数の負例候補が存在する．このように正例と負例の数が偏った状況では，学習される抽出規則の精度は低い．そこで，適切な数の負例に絞り込む必要がある．

本方式では正例を含む文書データ中で高頻度に出現する単語を頻度順に作業者に提示し，負例として適切なものを選択させる．これは，少ない負例候補の選択で多くの事例を得るためである．

2.3.2　規則の確信度

教師データにおいて正例および負例を分類する抽出規則の精度を抽出規則の確信度と呼ぶ．確信度は，教師データ中の NE を正例に分類した回数を P，負例に分類した回数を N とした場合，次式で表される．

$$\text{確信度} = \frac{P}{P+N}$$

例えば，次の抽出規則の確信度は 0.93 である．

「IF 前 1 単語=“ 担当 ” AND 後 1 単語=“ 見積 ” THEN 正例（製品名）」

すなわち，この規則が教師データにおいて 100 個の NE を分類した場合，そのうちの 93 個は正例，残り 7 個が負例であったことを表す．

2.3.3　NE 候補の確信度

抽出規則と同様に，抽出された NE 候補にも確信度が付加される．NE 候補の確信度は，抽出規則によりその候補が正例として抽出された回数を P，負例として抽出された回数を N とした場合，2.3.2節で述べた確信度の式により求められる．本稿では，「確信度の値が高い候補は NE である可能性が高い」という仮説に基づいて，確信度の高い順に候補を提示する．

3.　評価実験

3.1　実験方法

本方式を JAVA により実装し，NE の一つとして製品名を抽出する実験を行った．実験データとして営業日報データ（58,807 文書, 22MByte）を利用した．対象単語の前後の単語数 n は，予備実験の結果 3 とした．また作業は，筆者の一人が担当し，学習のサイクルは 3 回行った．

3.2　実験結果

表 1に各サイクルで入力した正例および負例の数，NE 抽出までの所要時間，抽出規則の数および規則により抽出した NE 候補の数をまとめる．正例と負例の数は異なり語数[1]である．また，正例の数，負例の数および抽出規則の数は累計である．NE 候補（製品名候補）の数はサイクル毎に新たに抽出した製品名候補の数を表す．第 1 サイクルの所要時間が第 2 サイクル以降よりも時間がかかるのは，文書データを C4.5 に入力可能なデータ形式に変換する処理が実行されるためである．

[1]文書中で，ある単語が何度用いられていてもこれを一語とし，全体で異なる単語がいくつあるかをかぞえた数．

表 2に各サイクルで生成された抽出規則の一例を示す．例えば，第 1 サイクルで生成された規則 1-1 は「後 1 単語が“ 紹介 ”であれば，0.95 の確信度で製品名である」ことを表す．実験では確信度が 0.80 未満の規則を削除した．表 2では規則 1-2, 2-2 が削除された．

表 3は第１サイクルの実行によって抽出された製品名候補の一部を示す．評価の○印は製品名，△印は製品カテゴリを表す語，×印は製品名以外の語であることを示す．表 3より，確信度が高い方が製品名が多いことが分かる．

また，第１サイクルで抽出した製品名候補を人手で評価した結果を表 4に示す．表 4は，確信度別に製品名が占める割合を示したものである．表 4より，確信度が高い候補に製品名が含まれる割合が高い．このことから，2.3.2節で述べた仮説どおり「確信度の値が高いと製品名である可能性が高い」ことが言える．

表 1　製品名抽出実験の入出力

サイクル	正例の数	正例として入力した単語	負例の数	負例として入力した単語	所要時間	抽出規則の数	NE候補の数
1	11	“JP1”, “HA8000”, “HITPHAMS”など	20	“予算”, “者”, “回答”など	約24時間（システム：約23時間，利用者：約1時間）	160	2724
2	1214	“SANRISE2000”, “HiRDB”など	1200	“デモ”, “データ”, “状況”など	約13時間（システム：約12時間，利用者：約1時間）	384	492
3	1504	“GEMPLANET”, “DF350” など	1500	“ヒアリング”, “表示”, “情報”など	約13時間（システム：約12時間，利用者：約1時間）	472	567

表 2　製品名の抽出規則（一例）

1	1-1	後1単語=“紹介”	正例(製品名)	0.95
	1-2	後1単語=“説明” and 前3単語目の品詞=“名詞” and 後3単語目の品詞=“名詞”	正例(製品名)	0.73
	1-3	後1単語=“申請”	負例	0.98
2	2-1	後1単語=“サポート” and 文頭	正例(製品名)	0.81
	2-2	後1単語=“検討” and 文末	正例(製品名)	0.77
	2-3	後1単語=“月”	負例	0.97
3	3-1	前1単語=“サーバー”	正例(製品名)	0.95
	3-2	前1単語=“1” and 後2単語目の品詞=“名詞”	正例(製品名)	0.89
	3-3	前1単語=“業務” and 後1単語=“紹介”	負例	0.80

表 3　抽出された製品名候補（一部）

製品名候補	確信度	出現頻度	評価	解説
ＨＩＴＰＨＡＭＳ	1.00	57	○	生産管理システム
Ｄｏｓ	1.00	47	△	製品のカテゴリ
ＤＦ３５０	1.00	18	○	ディスクアレイ装置
ＮＣＤ	1.00	12	×	譲渡性預金
ＨＩＴＡＣ	0.97	156	○	汎用機
…	…	…	…	…
ＳＡＮＲＩＳＥ	0.73	1292	○	ディスクアレイシステム
ＤＩＳＫ	0.71	245	×	ディスク
ＤＨＣＰ	0.70	20	×	プロトコル
Ｈｕｍａｎｉｍａｔｅ	0.68	64	○	人材管理システム
ＣＡＤ	0.68	327	△	製品のカテゴリ
ＭＲＰ	0.65	132	×	資材所要計画
…	…	…	…	…

表 4　確信度と製品名候補数の関係

	確信度		
	0.80以上	0.80未満かつ0.60以上	0.60未満かつ0.50以上
製品名の候補数	2307語	349語	68語
候補において製品名が占める割合	約7割	約4割	約2割

3.3　製品名の抽出精度

表 5は抽出精度を計算した結果を示す．抽出精度には次式で定義される F 値を採用した．

表 5　製品名の抽出精度のＦ値による評価

サイクル	候補数	適合率(Precision)	再現率(Recall)	抽出精度（F 値）
1	2724	0.44 (1214/2724)	0.61 (1214/1989)	0.51
2	492	0.46 (1504/3216)	0.75 (1504/1989)	0.57
3	567	0.40 (1535/3783)	0.77 (1535/1989)	0.53

$$\text{F 値} = \frac{(\beta^2+1)\times precision\times recall}{\beta^2\times precision+recall}$$

ここで *precision*（適合率）と *recall*（再現率）は，正解と一致するものの数を x としたとき，以下で定義される．

- 適合率：*precision* = x /抽出した NE の数
- 再現率：*recall* = x /正解リストにおける NE の数

ただし，正解リストとは事前に営業日報データに含まれる製品名 1,989 語を人手で収集したリストである．また，F 値は NE 候補の異なり語数をもとに計算し，適合率と再現率の相対的な重みを表すパラメータ β は 1.0 とした．

表 5に示した結果より，抽出精度（F 値）は第 2 サイクルの抽出において 0.51 から 0.57 に向上した．これは，第 1 サイクルよりも第 2 サイクルの方がより多くの正例を与えたことで，精度良く多くの製品名を抽出することができたと考えられる．また，第３サイクルでは，精度が低下するが，再現率が向上している．すなわち，正解リストにある製品名の約 77%を抽出することができた．

3.4　負例を与えることの効果

本方式では，正例と負例から抽出規則を学習し，過剰に汎化された規則による抽出誤りを防いでいる．表 2に示した規則 3-3 「IF 前 1 単語=“ 業務 ” and 後 1 単語=“ 紹介 ” THEN 負例」は，規則 1-1「IF 後 1 単語=“ 紹介 ” THEN 正例(製品名)」の条件部に「前 1 単語=“ 業務 ”」が加えられたより特殊な規則である．規則 3-3 を用いることで，その規則よりも一般的な規則 1-1 の抽出誤りを防ぐことができる．例えば，営業日報中に「…銀行に JP1 を紹介…」(…/銀行/JP1/紹介/…)や「…業務の詳細を紹介…」(…/業務/詳細 /紹介/…)といった文書が多数含まれる場合（括弧内は形態素解析後の単語列を表す），規則 1-1 だけでは，製品名“ JP1 ”以外に“ 詳細 ”といった NE 以外の単語を NE として抽出してしまう．これに対し，規則 3-3 を適用することで“ JP1 ”のみを NE として抽出することができる．

3.5　確信度の効果

本方式では，作業者が確信度という評価値を参考にすることで， 正しい NE 候補を効率良く判別できる．表 3に示したように，確信度の高い製品名候補はほぼ正解であった．作業者は確信度順に判別作業を進めることで，容易に多くの製品名を得ることができた．もし本方式において，作業者が確信度ではなく頻度順に NE 候補を判別する場合，効率的に製品名を抽出することは困難になる．例えば，表 3に示した抽出結果を頻度順に並べると，に示すように製品

名ではない候補が上位に現れ，容易に製品名を得られなくなる．本方式では確信度を用いることで，この問題を解消している．

表 6　頻度順に並べた製品名候補（一部）

製品名候補	確信度	出現頻度	評価
ＣＡＤ	0.68	327	△
ＤＩＳＫ	0.71	245	×
ＶｏＩＰ	0.89	147	×
ＭＲＰ	0.65	132	×
ＦｉｒｅＷａｌｌ	0.83	98	×
Ｈｕｍａｎｉｍａｔ	0.68	64	○
ＨＩＴＰＨＡＭＳ	1.00	57	○
…	…	…	…

3.6　本方式の限界

本方式では，NE と同義の単語を区別することができない．例えば，製品名“ Prius ”と一般名詞“ PC ”が，「…向けの Prius を紹介…」、「…向けの PC を紹介…」というように同じ文脈に出現する場合，本方式で採用したデータの属性，すなわち前後 n 単語の文字列，品詞および文字種を用いた規則（例えば，表 2に示した規則 1-1）では区別できない．このことから，上述の属性だけで，正確に抽出規則を学習することは困難である．

本実験で用いた営業日報は IT 関連の製品名が多数出現する．そのような製品名の多くは英数字からなる単語であり，形態素解析後でも１語の未知語として扱われる．そのため本方式では，単語の分割誤りを考慮せずに NE 抽出を行えた．しかしながら，実際にはこのようなケースは稀である．例えば，組織名“ 日立製作所中央研究所 ”は，「日立／製作所／中央／研究所」といったように，複数の形態素からなる．本方式では複数の形態素からなる NE を１語として扱えないので，NE の一部の形態素しか抽出することができない．

4.　関連研究

本稿と関連する研究に，1章で述べた宇津呂らの方式（以下，従来方式１）[7]や，磯崎の方式[6]（以下，従来方式２）がある．従来方式１の特徴は，最初に少量の教師データを与えるだけで，抽出規則を逐次的に学習して，対象文書から NE 候補を抽出する点である．学習の各サイクルでは，抽出した NE 候補全てを自動的に正例として用いる．これに対し本方式では，NE 候補を確信度順に人手で判別して，正しい NE を正例，NE 以外の単語を負例として，それらを再び学習に用いている．

従来方式２は，ブートストラップ法を採用していないが，本方式と同様に C4.5 を用いて NE 抽出を行う．この方式の特徴は，大量の教師データの他に NE に関するメタルールを事前に用意して，高精度の抽出規則を学習できる点である．メタルールは NE に関する特徴を記述した規則であり，抽出規則の精度向上に寄与している．これに対して，本方式は単に教師データだけを用いて抽出規則を逐次的に学習している．

5.　おわりに

本稿では，NE 抽出規則である決定木を逐次的に学習するステップにおいて，確信度に基づく NE 候補の判別作業を導入した NE 抽出方式を提案し，営業日報を用いた実験を通じて有効性を確認した．

今後の課題として，本方式による NE 候補判別にかかる人手のコストを抑えることを検討する．また，形態素解析により複数の形態素に区切られた NE を一つの NE として抽出できるように本方式を改良する．改良した本方式を用いて製品名以外の NE を抽出する実験を行い，その結果を考察する．

6.　参考文献

[1] 関根聡：固有表現から専門用語，言語処理学会第 10 回年次大会併設ワークショップ「固有表現と専門用語」発表論文集，pp.1-4 (2004).

[2] R. Grishman and B. Sundheim : Message Understanding Conference - 6: A Brief History, Proceedings of the 16 th International Conference on Computational Linguistics (COLING-96), pp. 466-471, (1996).

[3] S. Sekine and H. Isahara : IREX:IR and IE evaluation-based project in Japanese, Proceedings of the 2nd International Conference on Language Resource and Evaluation (2000).

[4] IREX 実行委員会：
http://www.csl.sony.co.jp/person/sekiNE/IREX/NE/
(1999).

[5] 山田寛泰，工藤拓，松本裕：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).

[6] 磯崎秀樹：メタルールと決定木学習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.5，pp.1481-1491 (2002).

[7] 宇津呂武仁，颯々野学：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，情報処理学会研究報告，2000-NL-139，pp.9-16 (2000).

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

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

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

[11] J.R. Quinlan : C4.5 : Programs for Machine Learning, Morgan Kaufmann (1993).