固有表現抽出ツールＮＥｘＴの精緻化とユーザビリティの向上

渡辺　一郎†　桝井　文人†　福本　淳一‡

†三重大学工学部　‡立命館大学理工学部

１　はじめに

固有表現抽出は，文書中の人名，地名などの固有名詞や金額，割合といった数量表現を抽出，タグ付けを行う情報抽出技術である．MUC[8]やIREX[9]で固有表現抽出タスクが設定され，多くの研究機関が参加した．その後も盛んに研究が行われ，パターン駆動型の他に SVM[3]や最大エントロピー法 [4]や複数の抽出モデルを混合した手法[5]などが提案されている．NExT（Named Entity Extraction Tool）は，パターン駆動型汎用固有表現抽出ツールとして公開されている[1][2]．そして，NExT の探索手法，パターン辞書，抽出精度に問題があった．

本論文では，NExT をより高度な固有表現抽出ツールとするため，これらの問題を解決するいくつかの手法を提案し精緻化とユーザビリティを向上させる．また．固有表現抽出タイプを従来の７種類から２００種類に拡張したので報告する．

以下，２章で NExT の概要と現在の問題点を述べ，３章，４章で提案手法である文末からの探索，固有表現キー・抽出規則の精緻化を説明し，５章で固有表現抽出タイプの拡張とシステム変更，６章で単語反復度に基づくキー獲得，７章で実験と考察を行い，８章でまとめを行う．

表１　パターン辞書

固有表現抽出タイプ	固有表現キー例
人名	さん，首相，教授
地名	氏，半島，地方
組織名	病院，空港，協会

２　NExT の概要と問題点

NExT は，ChaSen，Juman の形態素解析[6] [7]処理済みのテキストを入力とし，様々な形式で出力可能な，柔軟なインターフェースと，簡単な操作性，及び，目的に応じて手軽に拡張や改変が可能な拡張性を持つ．

しかしながら，現在 NExT には以下のような問題点がある．

（１）固有表現キー処理の探索は，文頭から行っており固有表現キー検出後，固有表現の範囲を決定する際に，探索方向とは逆方向に戻るため抽出処理効率が悪い．

（２）登録されているパターン（固有表現キー）では，固有表現だけを抽出するには不十分な場合があり，誤抽出，過抽出の原因となっている．

（３）固有表現キーの付属しない場合，固有表現を抽出できない．

（４）システムで複数の処理が同時に行なわれており，処理の追加，削除が困難である．

３　文末からの探索

本章では前述した問題を解決するための手法を提案する．固有表現キーが多くの場合，固有表現の終端に位置する事を利用し，照合処理を文末から行う．これにより，固有表現の範囲を決定する際の無駄なバックトラックを減らすことができ，処理速度の向上につながる（図１）．また，１つの固有表現に複数の固有表現キーが含まれる場合に，一度の探索で全体を抽出することが可能となる．例えば，従来手法では，『サマワ市評議会』を抽出する際には，まず，固有表現キー“市”に一致し“サマワ市”が抽出され，次に“会”によって全体を抽出していた．しかし，提案手法では，後方から探索を行うため，はじめに“会”に一致し，そのまま文頭へ向けて固有表現スコープの探索を行い，“市”は無視される．その結果，バックトラックを行わずに，“サマワ市評議会”を抽出できる．

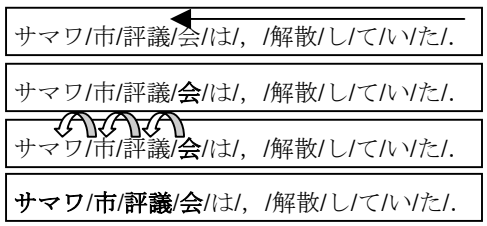

図１　文末からの探索

４　固有表現キー・抽出規則の精緻化

NExT の精度は，固有表現キーの質に影響される．従来登録されている固有表現キーは，１形態素で構成されており，固有表現だけを抽出することは困難であった．例えば，従来の固有表現キー“長官”で，『福田官房長官』を抽出すると，『福田官房』までが人名とされ，誤抽出になる．提案手法では，複数形態素固有表現キー（表２）を利用し，『福田』が正しく抽出できる．

表２　複数形態素固有表現キーの例

抽出タイプ	固有表現キー
人名	官房/長官，総理/大臣，名誉/教授
地名	自動車/道，国立/公園
組織名	事業/団，同好/会，委員/会

また，これまで抽出規則は，システムに統合され固有表現抽出タイプ毎に固定であった．しかし，固有表現キー毎に，抽出規則を切り替えて使用することで，同一固有表現キーであってもタイプを分けて抽出する事ができる．例えば，組織を抽出する固有表現キー“社”を使用しこれまでの抽出規則を適用することで,『オラクル社』のような組織名が抽出できる．さらに,“社”に「名詞-数詞」が，接続する場合には,『４社』のような組織数を抽出できる．

５　システムの拡張と変更

NExT は，固有表現キーが存在しない固有表現に対して，形態素処理以外の抽出方法が無い．特に，人名，組織名は，固有表現キー付属しない場合がある．また，Web 文書，専門書などでは，専門用語や固有表現の省略語が利用される．このような固有表現に対応するために，辞書に登録された固有表現を単純マッチングで抽出する機能を実装した．さらに，今後の固有表現の利用範囲を広げるために，関根の拡張固有表現階層に基づき，従来７種類だった固有表現抽出タイプを，２００種類に拡張した．これは，表３のように，これまでの７種類のタイプの中を細かく分類するだけでなく，さらに，７種類以外で抽出しなかった新たなタイプを追加し抽出する．

また，従来システムは，規則と処理が一体化しており，メンテナンス性が悪かった．そのため，規則と処理を完全に分離し，処理もそれぞれ独立したモジュールにすることで, 処理の追加, 変更, 削除が容易にできるようにした．

６　単語反復度に基づくキー獲得

NExT のパターンである固有表現キーはこれまで人手で獲得されていた．固有表現抽出タイプを拡張したことで，これまで以上に多くの固有表現キーが必要である．

日本語においても単語反復度がキーワード特定の特徴量となることが報告されている[10]．抽出タイプ毎に分類された固有表現から, SuffixArray を作成し一定頻度以上の Suffix を固有表現キー候補とし獲得する．この固有表現キー候補を用いて，

表３　拡張タイプの例

地名	都道府県名・市町村名など
組織名	企業名・軍隊名・協会名など
イベント名	事件事故名・自然災害名など
寸法表現	長さ・面積・重量・速度など

タイプ毎に分類された固有表現に対して，固有表現抽出を行い，抽出可能であった固有表現キー候補を，新たな固有表現キーとして獲得する．

この手法を用い抽出タイプ毎に分類された 56757 個の固有表現に対して行い 1719 個の固有表現キーを獲得した．

７　実験と考察

実験には，IREX の NE_DRYRUN[9]の毎日新聞９４年版の３４記事と MET(Multilingual Entity Task)[8]を使用し，従来手法と提案手法の比較を行った．提案手法には，416 個の複数形態素固有表現キーを含む辞書を使用し，従来手法には，複数形態素固有表現キーを含む辞書より１形態素の固有表現辞書を作成した．また，単純マッチングには，人手で作成した 7204 個のマッチング辞書を使用した．さらに，IREX のタスク定義に対しては，単語反復度に基づく固有表現キー獲得手法を用い獲得した固有表現キーと，人手で獲得した 13023 個のマッチング辞書を使用した場合に対しても行った．

実験結果を，表４，表５に示す． IREX に基づくタスク定義の実験で再現率が 12.0%，適合率 13.4%で，Ｆ値で 11.6 向上し，MET に基づくタスク定義の実験で，再現率が 14.4%，適合率が 16.6%，Ｆ値が 8.8 向上している．これは，抽出不可能だった『NASDA』や『石播』といった省略語，『地球』『火星』などの地名が単純マッチング機能により抽出できたためである．

また，従来手法では,『宇宙局』と『米航空宇宙局』,『対策本部』と『事故対策本部』の両方を抽出しているが, 提案手法では,『米航空宇宙局』『事故対策本部』のみを抽出し過抽出を防ぐことができた．また,『クエール副大統領』を『クエール副』と抽出していたが『クエール』と正しく抽出され誤抽出が減少した．一方，時間表現においては従来手法を下回った．これは,『一月四日朝』などの表現を分割せずに，まとめて出力するようにしたためである．また，従来の NExT は，IREX タスク定義の ARTIFACT（人工物）の抽出に対応していなかったが，提案手法を用いることで抽出可能となり，他手法と同等の評価ができる．限定ドメ

表４　固有表現抽出結果

	IREX に基づく評価						MET に基づく評価					
	従来手法			提案手法			従来手法			提案手法		
	再現率	適合率	F 値	再現率	適合率	F 値	再現率	適合率	F 値	再現率	適合率	F 値
組織	58.9%	63.6%		62.6%	77.9%		67.7%	58.4%		88.7%	84.6%	
人名	76.9%	63.4%		88.8%	77.7%		83.1%	85.0%		96.4%	73.7%	
地名	67.7%	63.7%		79.7%	86.0%		66.9%	96.2%		89.8%	88.5%	
人工物				23.8%	54.5%							
日付	78,0%	80.2%	69.8	91.7%	91.7%	78.1	93.6%	82.0%	76.9	97.6%	95.7%	85.7
時間	78.3%	64.3%		69.6%	69.6%		91.3%	100.0%		77.4%	78.1%	
金額	93.9%	93.9%		93.9%	93.9%		98.6%	100.0%		100.0%	100.0%	
割合	66.7%	100.0%		66.7%	100.0%		77.3%	77.9%		77.3%	100.0%	
全て	70.2%	67.4%		75.9%	79.6%		73.7%	68.5%		88.1%	85.1%	

インの性能は，他手法の[3，4，5]と同程度の性能を示した．固有表現キーとマッチング辞書の登録数を増やした場合には，一般ドメインでも，同等の性能を示した．

処理速度は，複数固有表現キー辞書の利用によるパターンの増加と，システムに統合されていた抽出規則を分離したことにより辞書読み込みのオーバーヘッドが増加した．しかし，探索方法や実装方法の変更によって 1 文書当たりにかかる処理時間は，従来と同程度で行うことができる．

8　おわりに

本論文では，探索手法の変更と固有表現キー・抽出規則の精緻化，単純マッチングによる抽出機能を提案し，実験によりその有効性を確認した．また，今後の応用範囲を広げるため，固有表現抽出タイプを拡張し，さらにシステム構成を変更することで抽出精度とユーザビリティを向上させた．

今後の課題として，さらに，『〇分の△』やメールアドレスのように固有表現キーが間に入っている場合や，『石破防衛庁長官』のような人名の中に組織名が含まれている固有表現が正しく抽出できない．これらに対応する処理の追加が必要である．

表５　IREX に基づく，辞書規模による結果

	従来の辞書			大規模辞書		
	再現率	適合率	F 値	再現率	適合率	F 値
組織	62.6%	77.9%		76.2%	82.7%	
人名	88.8%	77.7%		92.3%	82.1%	
地名	79.7%	86.0%		77.6%	80.5%	
人工物	23.8%	54.5%		28.6%	40.0%	
日付	91.7%	91.7%	78.1	91.7%	91.7%	81.5
時間	69.6%	69.6%		69.6%	69.6%	
金額	93.9%	93.9%		93.9%	93.9%	
割合	66.7%	100.0%		66.7%	100.0%	
全て	75.9%	79.6%		82.2%	80.8%	

参考文献

[1] 桝井文人，鈴木伸哉，福本淳一: "テキスト処理のための固有表現抽出ツール NExT の開発"，第 8 回言語処理学会年次大会発表論文集, pp.176-179, 2002.

[2] 渡辺一郎，桝井文人，河合敦夫: "固有表現抽出の性能と分野依存性の検証"，平成 15 年度電気関係学会東海支部連合大会公演論文集, 536, 2003.10.

[3] 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，情報処理学会誌，Vol. 43，No. 1,pp.43-53, 2002.

[4] 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. 自薦言語処理, Vol. 7, No. 2, pp.63-90m 2000.

[5] 宇津呂武仁，颯々野学，内元清貴．正誤判別規則学習を用いた複数の日本語固有表現抽出システムの出力の混合．四川言語処理，Vol9, No. 1, pp.65-100, 2002.

[6] 松本裕治，北内啓，山下達雄，松田寛，高岡一馬，浅原正幸：形態素解析システム茶筅 version 2.3.3 使用説明書, 2003.

[7] 黒橋禎夫，長尾真：日本語形態素解析システム JUMAN 使用説明書 version3.61，京都大学大学院工学研究科.

[8] DARPA. Proceedings of the Tipstar Text Program PhaseⅢ18 month workshop. DARPA, 1998.

[9] IREX 実行委員会（編）．IREX ワークショップ予稿集

[10] 武田善行，梅村恭司, "キーワード抽出を実現する文書頻度分析"，計量国語学，Vol.23，No.2，pp.27--32, September 2001.