E-009

長い複合名詞の構造的な固有表現認識

Structural Named Entity Recognition in Long Compound Noun

船山 弘孝　柴田 知秀　黒橋 禎夫
Hirotaka Funayama　Tomohide Shibata　Sadao Kurohashi

1　はじめに

固有表現認識 (Named Entity Recognition) とは、テキスト中から人名、組織名、地名などの固有表現の認識を行なう処理である。固有表現は常に新しいものが生まれ続けており、またテキストは人手の処理能力以上に大量に存在するため、その全てを手作業により辞書に列挙することは難しい。そのため、Web や新聞記事などの大規模コーパスを利用して固有表現認識に関する多くの研究が行われており、Support Vector Machine (SVM) や Conditional Random Field (CRF) などを用いた手法が提案されている。日本語において、これらの研究が対象としている固有表現のタイプは、IREX(Information Retrieve and Extraction Exercise) で定義されている組織名, 人名, 地名, 固有物名, 日付表現, 時間表現, 金額表現, 割合表現の 8 種類である。

一般に固有表現認識では系列ラベリングを用いて行われる場合が多い。系列ラベリングによる手法は高い精度が報告されているが、例えば Sasano らの手法 [1] は 5 形態素以上の長い固有表現に対しては F 値 80 程度の精度で十分とは言えない。そこで本研究では、長い固有表現も精度よく認識するために、あらゆる複合名詞のラベルを SVM を用いて推定し、推定されたラベルから CKY 法を用いて構造的にラベルを決定することにより固有表現認識を行う。

2　関連研究

一般に固有表現認識は、系列ラベリング問題で定式化される。系列ラベリング問題とは、データの系列に対してラベル付けを行う問題のことである。固有表現解析では、形態素列 (文) の中から固有表現である形態素列を認識するが、これは各形態素に対して、1 形態素の固有表現 (S)、複数形態素の固有表現の最初 (B)、途中 (I)、最後 (E)、固有表現ではない (O) のいずれかのラベルを付与する問題と考えることができる[1]。

複数の固有表現について考える場合は、各固有表現に関して S,B,I,E を組み合わせたラベルについて考える。表 1 にこのようなラベル付けによる固有表現認識の例を示す。

このようなラベル付けは、コーパスに対して表 1 のような形式で固有表現認識の正解データを付与し、そこから機械学習によってラベル付けの判定器を学習することで実現できる。このとき学習の素性としては、通常、ラベル付けする形態素の前後 5 形態素の文字列、品詞、文字種などが用いられる。例えば、表 1 で “中田” のラベル付けをする場合は “年” から “は” までの文字列、品詞、文字種などを用いる。学習アルゴリズムとしては、SVM,CRF などを用いる方法が提案されている。

表 1: 系列ラベリングによる固有表現認識

形態素 (文)	ラベル	素性として用いる文字列		
		文字列	品詞	文字種
2001	B-DATE	2001	数詞	数字
年	I-DATE	年	助数詞	漢字
夏	E-DATE	夏	時相名詞	漢字
中田	B-PSN	中田	人名	漢字
英寿	E-PSN	英寿	人名	漢字
は	O	は	副助詞	平仮名
ローマ	S-ORG	ローマ	地名	片仮名
から	O	から	格助詞	平仮名
移籍	O	移籍	サ変名詞	漢字
。	O	。	句点	記号

日本語の固有表現認識では、新聞記事 1 万文に対して、約 2 万個の固有表現が付与された CRL 固有表現データが用いられる場合が多く、このデータに対して F 値 89 程度の解析が実現されている [1, 4, 5, 6]。

3　提案手法

系列ラベリングによる固有表現認識では、形態素や文字を単位として B-PSN,I-PSN,E-PSN などをタグ付けし、その結果をまとめて PERSON を認識するが、これは我々の直感的な理解とは若干異なる。例として

マイケル・マカリー国務省首席報道官を ...　(1)

という文から “マイケル・マカリー”(PERSON) という固有表現を認識する場合を考える。この例において例えば “マカリー” という形態素のラベルを推定しようとする際に系列ラベリングによる手法では “マイケル”〜“省” の 5 形態素の情報を用いて “マカリー” が PERSON の末尾 であると推定するが、実際我々が認識を行う際には “報道官” に “マイケル・マカリー” という複合名詞がかかっているために “マイケル・マカリー” 全体で PERSON であるというように判断する。

そこで本研究では、CKY 法を用いて構造的に固有表現認識を行う手法を提案する。以下に提案手法につい

京都大学大学院情報学研究科

[1]このようなラベル付けの手法を SE 法 [2] という。これ以外にも I,O,B のみを用いた IOB1,IOB2、I,O,E のみを用いた IOE1,IOE2 法 [3] などがある。

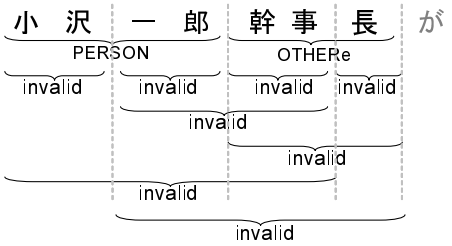


図 1: 文節 “小沢一郎幹事長が” に対するラベルの与え方

て、学習時と解析時の 2 つに分けて簡単な流れを示す。

3.1 学習

まず、文節の前後の機能語を削除する。例えば “小沢一郎幹事長が” という文節なら最後の助詞 “が” を削除して “小沢一郎幹事長” とし、この単位に対して学習を行う。ただし、括弧でくくられた文節は機能語を削除しない。これは、括弧でくくられた表現は機能語も含めて固有表現になりやすいためである。

次にこのよう前後の機能語を削除した文節に対して以下のような形式でラベルを付与する。このとき考えるラベルは IREX で定義された LOCATION、ORGANIZATION、PERSON、ARTIFACT、DATE、TIME、PERCENT、MONEY の 8 種類とそれ以外の表現に対する OTHER_{sigle}、OTHER_{begin}、$\mathrm{OTHER}_{intermediate}$、$\mathrm{OTHER}_{end}$、invalid の 5 種類を加えた計 13 種類である。図 1 に “小沢一郎幹事長が” という文節に関するラベル付けの例を示す。

IREX で定義された 8 種類のラベルに関しては、正解データを元にラベルを付与する。図 1 の例では “小沢一郎” が PERSON となる (以降このようなラベル付けの単位をチャンクと呼ぶ)。

また、OTHER_{single} は文節全体がラベルなしの場合にその文節全体に与えるラベル、OTHER_{begin}、$\mathrm{OTHER}_{intermediate}$、$\mathrm{OTHER}_{end}$ はそれぞれ文節先頭、文節途中、文節末尾がラベル無しの場合に該当するチャンクに与えるラベルである。図 1 の例では “幹事長” が OTHER_{end}[2]となる。

IREX で定義された 8 種類の固有表現でもなく OTHER の 4 種類でもないものには invalid というラベルを与える。図 1 の例では、上記以外の “小沢”, “小沢一郎幹事”, “小沢一郎幹事長”, “一郎”, “一郎幹事”, “一郎幹事長”, “幹事”, “長” がこれに当たる。このラベルを考えることで “小沢” や “一郎” というチャンクに対して PERSON とラベル付けせず、“小沢一郎” というチャンクに対して PERSON とラベル付けするようなモデルを構築できる。このようにしてラベル付けされたチャンクそれぞれに対して表 2 のような素性を考える。

表 2 において (1)−(6) に関してはチャンクの先頭または末尾であるという素性も付与した。例えば “小沢一郎幹事長” というチャンクには、先頭が “小沢” で「人名」、末尾が “長” で「人名末尾」のような素性が付与される。なお、先頭と末尾以外の位置情報は考慮していない。また (7),(8) の素性に関しては、その文節から生成される全てのチャンクに付与した。

表 2: 考える素性

(1) チャンク内にある文字種が含まれるかどうかや文字種の順番
  - 文字種としては漢字、平仮名、片仮名、(漢) 数字、アルファベットの 5 種類とする

(2) 考えているチャンクに含まれる形態素数

(3) チャンクが文節先頭から始まっているかや文節末尾で終わっているかというチャンクの文節内における位置に関する情報

(4) JUMAN によって付与される品詞・品詞細分類・カテゴリなどの情報

(5) KNP によってチャンクを含む文節、チャンクに含まれる形態素に付与される Feature [4]

(6) 文字列自体
  - 文節主辞、形態素、助詞、係り先の主辞など

(7) 用言格フレームの固有表現に関する頻度情報 [7]
  - 例えば “小沢一郎幹事長が会見した。” という文で、“会見する” という用言のガ格に「PERSON:0.245」などの頻度情報があった場合は 0.245 を “小沢一郎幹事長” から生成される全てのチャンクに対して与える

(8) 括弧に関する以下 2 つの素性
  - 括弧内のチャンクであるという素性
  - KNP により付与される括弧表現と同格である表現がある場合にはその同格表現の文字列そのもの

(9) 前の文節の最後のチャンクのラベル
  - この素性は文節の先頭が数詞でありかつ前の文節の最後が名詞であるというチャンクにのみ付与する

以上の各チャンクに対するラベル-素性から SVM を用いてモデルを作成する。SVM は 2 値分類器であるため one vs rest 法を用いて多値分類器に拡張する。one vs rest 法は、$\{C_1, C_2, \ldots, C_n\}$ への多値分類の問題を C_k かそれ以外かという 2 値分類に分解する手法である。ここでは 13 種類のラベルを考えているので、SVM を用いて 13 個のモデルを作成した。

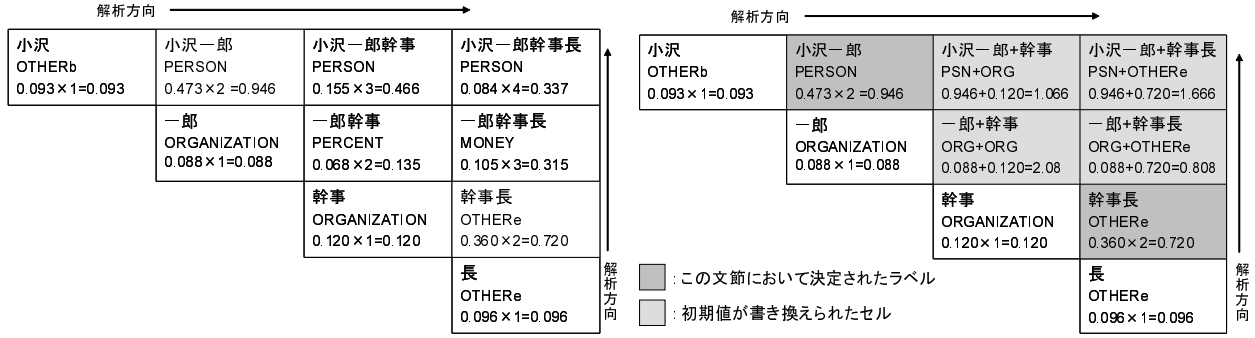
3.2 解析

解析時に考える単位は学習時と同様、文節内の全ての連続する形態素とする。例えば “小沢一郎幹事長が” という文節なら、まず文節の前後から機能語 “が” を削除した後、“小沢”, “小沢一郎”, “小沢一郎幹事”, “小沢一郎幹事長”, “一郎”, “一郎幹事”, “一郎幹事長”, “幹事”, “幹事長”, “長” の各候補についてそれぞれラベルを推定したのち、以下に述べる手法で文節単位のラベルを決定する。

まず、各チャンクについて SVM の出力のスコアが最も高かったラベルとそのスコアを利用し、CKY 法を用いてどのパスが正しいかを決定する。ここで SVM

[2]各 OTHER は “幹事”,“長” のように形態素単位にラベル付けを行わず、文節内で最長となるようにする。

[4]“大分” が地名 “おおいた” と副詞 “だいぶ” の 2 つの解釈があるように 2 つ以上の解釈がある場合はその曖昧性も考慮している。



a:初期状態　　b:CKY 法適用後

図 2: 文節 “小沢一郎幹事長” に対して CKY 法を適用した例

の出力のスコアはシグモイド関数 $\frac{1}{1+exp(-\beta x)}$ を用いて変換し、さらに各ラベルのスコアの和が 1 になるように正規化する。この正規化したスコアから現在考えているチャンクのスコア

$$Score = \sum_{k=1}^{n} m_k S_k \qquad (2)$$

を計算する。n は現在考えているチャンク内の全ラベル数、S_k は各ラベルに対する SVM の出力のスコア、m_k は k 番目のラベルに含まれる形態素数である。ここで、ラベルが invalid であると推定されたチャンクに対しては 2 位のスコアとラベルのペアをそのチャンクに対するスコアとラベルとする。このようにして得られたスコアから図 2 のように CKY 法を適用して最も高いスコアが与えられたパスを決定しその文節のラベルとする。構文解析の際に用いられる一般的な CKY 法と異なるのは、初期状態で対角線上以外の各セルにラベルとスコアのペアが入っており、各セルのラベルを決定する際に初期状態のスコアとの比較も行う点である。また、CKY 法でパスを選択する際には OTHER - OTHER の隣接を許さないなどの制約条件を与えている。

そして最後に複数文節にまたがるラベルがあればそれを接合し、最終的な出力とする。

4 実験

CRL 固有表現データに対して学習を行い、5 分割交差検定を行った。CRL 固有表現データでは、毎日新聞 95 年度版 1,174 記事、10,718 文に対して IREX で定義された 8 種類の固有表現がタグ付けされている。また、人手でタグ付けが困難であると判断された表現には OPTIONAL のタグが付与されているが、このような表現は学習には用いず[5]、評価する際も対象外とした。

また本実験を行うにあたり、形態素解析器として JUMAN[6]、構文解析器として KNP[7]を用いた。また SVM のカーネル関数には 2 次の多項式カーネルを用い、シグモイド関数の β を 1 とした。実験結果を表 4 に示す。全体の抽出精度は F 値で 87.10 であった。

表 4: 固有表現認識の結果

	Rec.	Pre.
ORGANZATION	78.56 (2888/3676)	86.31 (2288/3346)
PERSON	83.91 (3222/3840)	88.76 (3222/3630)
LOCATION	89.31 (4879/5463)	90.52 (4879/5390)
ARTIFACT	42.97 (321/ 747)	63.31 (321/ 507)
DATE	92.99 (3317/3567)	92.71 (3317/3578)
TIME	88.25 (443/ 502)	87.03 (443/ 509)
MONEY	94.62 (369/ 390)	96.85 (369/ 381)
PERCENT	92.28 (454/ 492)	95.58 (454/ 475)
ALL-SLOT	85.09	89.21
F-measure		87.10

5 考察

5.1 先行研究との比較

先行研究との比較を表 3 に示す。先行研究の中で高い精度が報告されているもののうち、風間らの手法や福島らの手法は wikipedia やウェブテキストなどの外部リソースから獲得した知識、Sasano らの手法では本手法では用いていないキャッシュや共参照などから得られる情報などを用いて学習している。そのため、本手法においてもこのような知識を用いることにより、さらに精度が向上すると思われる。

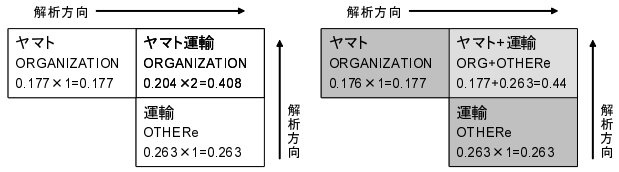
また本手法では、Sasano らの手法に比べて長い単位の固有表現について正しく認識できたものが多く見られた。例えば、“欧州通常戦力削減条約” という文節において Sasano らの手法では “欧州” が LOCATION とラベル付けされるが、本手法では “欧州通常戦力削減条約” が ARTIFACT と正しくラベル付けされる。

また “外国人登録法違反の” という文節において Sasano らの手法では “登録法” が ARTIFACT とラベル付けされるが、本手法では “外国人登録法” が ARTIFACT と正しくラベル付けされる。 Sasano らの手法はラベル付けを行う際に文節主辞の情報を用いているが、この例では主辞が “違反” であるため “外国” のラベルを推定する際に有用と思われる “法” の情報を用いていない。それに対して本手法では “外国人登録法” のチャンクを推定する際には “法” から得られる情報を用いることもできるため正しく推定できた思われる。

[5]ただし 4 種類の OTHER や invalid の学習には用いた。
[6]http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html
[7]http://nlp.kuee.kyoto-u.ac.jp/nl-resource/knp.html

表 3: 先行研究との比較

	CRL 交差検定	解析単位	その他の素性
福島ら 2008[4]	89.29	文字	ウェブテキスト
風間ら 2008[5]	88.93	文字	wikipedia, ウェブテキスト
Sasano and Kurohashi2008[1]	89.40	形態素	構造的情報
Asahara and Matsumoto2004[8]	87.21	文字	
中野ら 2003[6]	89.03	形態素	文節素性
磯崎ら 2003[9]	86.77	形態素	
提案手法	87.10	文節	



a:初期状態　　b:CKY 法適用後

図 3: エラー分析の例

5.2 エラー分析

本稿では CKY 法を用いて文節単位のラベルを決定する手法を提案したが、正しくラベル付けが行われていたチャンクが CKY 法を適用する過程で誤ったラベル付けが採用されて最終的な出力に反映されなかったものが存在した。例えば “ヤマト運輸は” という文節において “ヤマト運輸” に ORGANIZATION とラベル付けされるのが正解であるが、実際は図 3b のように “ヤマト” に ORGANIZATION とラベル付けされる。しかし、同図 a において “ヤマト運輸” という単位で見ると ORGANIZATION と正しくラベル付けされているにも関わらずそのラベル付けが採用されていないことが分かる。

このような誤ったラベル付けの比率を減らす方法としてシグモイド関数の β の値の調整、CKY 法適用のアルゴリズムや各チャンクに対する素性の工夫などが考えられる。

6 まとめと今後の課題

固有表現認識では系列ラベリングを用いて行われる場合が多いが、本研究では、あらゆる複合名詞のラベルを SVM を用いて推定し、推定されたラベルから CKY 法を用いて構造的にラベルを決定することにより固有表現認識を行い、87.10 の精度で固有表現を抽出できた。

今後は今回作成した固有表現に関するモデルを河原らによる構文・格解析統合モデル [10] へ組み込むことを目指す。河原らによる構文解析は、入力文がとりうる全ての構文構造に対して確率的格解析を行い、最も確率値の高い格解析結果をもつ構文構造を出力するモデルである。したがって CKY 法で文節のラベルを決定する際に用いた値に何らかの変換を施し、それを確率的値として用いることで、構文解析との統合が実現できると考えられる。

また、固有表現認識は辞書のなどの外部リソースを用いることで精度が向上することが報告されている [4, 5]。そこで Web 上の箇条書きや表から取得した同位語 [11]、Wikipedia から取得した知識などを利用することで抽出精度の向上が期待できる。文字単位でチャンキングを行う風間らの手法は Wikipedia から構築した上位語辞書から得られる情報を各文字に対して与えているが、本手法ではこの情報を 1 つのチャンクに対してのみ与えることが可能である。つまり、例えば図 1 の例において “小沢一郎” が “政治家” であるという wikipedia から得られた情報を使う際に、文字単位による手法では “小”,“沢”,“一”,“郎” の 4 つに対して分割して入れる必要がある。それに対して本手法では “小沢一郎” という 1 つのチャンクに対してこのような素性を容易に入れることができる。

参考文献

[1] Ryohei Sasano and Sadao Kurohashi, Japanese named entity recognition using structural natural language processing, *Third International Joint Conference on Natural Language Processing*, (2008), pp. 607–612.

[2] S.Sekine, R.Grishman, and H.Shinnou, A decision tree method for finding and classifying names in japanese texts, *6th Workshop on Very Large Corpora (WVLC-6)*, (1998).

[3] E. Tjong Kim Sang and J. Veenstra, Representing text chunks, *EACL '99*, (1999), pp. 173–179.

[4] 福島健一, 鍛冶伸裕, 喜連川優, 日本語固有表現に置ける超大規模ウェブテキストの利用, *DEWS2008*, (2008).

[5] 風間淳一, 鳥澤健太郎, Web 上の資源から構築した複数の固有表現辞書を用いた日本語固有表現認識, 言語処理学会 第 14 回年次大会発表論文集, (2008), pp. 813–816.

[6] 中野圭吾, 平井有三, 日本語固有表現抽出における文節情報の利用, 情報処理学会自然言語処理研究会 2003-NL-156-2, (2003), pp. 7–14.

[7] 笹野遼平, 河原大輔, 黒橋禎夫, コーパスサイズの拡大および用例の汎化による格フレームのカバレッジの改善, 言語処理学会 第 14 回年次大会発表論文集, (2008), pp. 528–531.

[8] Asahara Masayuki and Yuji Matsumoto, Japanese named entity extraction with redundant morphological analysis, *Proc. of HLT-NAACL 2003*, (2003), pp. 8–15.

[9] 磯崎秀樹, 加沢秀人, 固有表現抽出のための SVM の高速化, 情報処理学会論文誌, Vol. 44, No. 3, (2003), pp. 970–979.

[10] 河原大輔, 黒橋禎夫, 自動構築した大規模格フレームに基づく構文・格解析の統合的モデル, 自然言語処理, Vol. 14, No. 3, (2007).

[11] Keiji Shinzato and Kentaro Torisawa, Extracting hyponyms of prespecified hypernyms from itemization and headings un web documents, *COLING 2004*, (2004).