RE-002

二段階の機械学習を用いたボトムアップ型の固有表現認識

Bottom-up Named Entity Recognition using two-stage Machine Learning Method

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,NER) 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、情報抽出 (IE) や質問応答 (QA) など様々な高次タスクに利用されている。近年、人手でタグ付けされた正解データから Support Vector Machine[1] や Conditional Random Field[2] などの機械学習を用いて固有表現認識を行う手法が提案されている [3]。

一般に固有表現認識は、系列ラベリングという手法を用いて行われることが多い。例えば形態素単位に対して S-PERSON(1 形態素からなる人名)、B-PERSON(人名先頭)、I-PERSON(人名途中)、E-PERSON(人名末尾) などをタグ付けし、結果をまとめて PERSON(人名) などと認識する。

しかし、各形態素のラベルを推定する際は、例えば自分と前後 2 形態素の計 5 形態素程度の局所的な情報を用いるため長い複合名詞の解析が誤る場合がある。これを以下の例文を用いて考える。(「/」は形態素区切りを表す。)

(1) 帰国/した/風間/さん/は ...

(2) 発動/した/信用/組合/救済/銀行/の/設立/も

(3) 私文書/偽造/と/外国/人/登録/法/違反/の/疑い/で

(1) では接尾辞 “さん” や “帰国した” などの情報を用いることで、“風間” が S-PERSON であると認識することができる。

一方、(2) において、“信用” が ORGANIZATION の先頭 (B-ORGANIZATION) であることを推定する際には “発動”～“救済’ の情報を用いるだけであり、3 形態素離れた “銀行” の情報を利用することはできない。このような問題に対応するため、中野ら [4] や笹野ら [5] は、ラベルを推定する際に文節主辞の情報を用いている。

しかし、このような方法は (3) のように参照したい形態素が文節主辞でない場合は機能しない。この例において、“外国” が ARTIFACT の先頭 (B-ARTIFACT) であると推定する際には 文節主辞の “違反” ではなくその前の “法” が重要である。

そこで我々は複合名詞をラベル付けの単位 (以降このラベル付けの単位をチャンクと呼ぶ) とし、そこからボトムアップにラベルを決定することで固有表現認識を行う手法を提案する。本手法ではまず、文節内で考えられるあらゆるチャンクに対して機械学習によってラベルを推定する。あらゆるチャンクに対してラベル付けを行うので、上の例では “外国人登録法” という単位に対してもラベル付けを行うことができ、その際に “法” から得られる情報も利用することができる。次に、構文解析で用いられる CKY 法のようにボトムアップに解釈を決定する。その際、“外国人登録法”(ARTIFACT) と “違反”(OTHER) の組み合わせがよいのか、“外国人”(PERSON) と “登録法違反”(OTHER) の組み合わせがよいかなどを決定する。CRL 固有表現データを用いた実験を行ったところ F 値 89.67 となり従来研究よりも高い精度を達成した。

本論文の構成は以下のとおりである。2 章ではまず関連研究として固有表現認識の分野で一般的に用いられる、系列ラベリングについて説明する。次に 3 章では我々の提案手法の概要を、4 章では機械学習に用いるモデル作成とその素性に関する説明を行い、5 章では実際の解析について説明する。6 章では実験・結果、7 章ではその実験結果の考察を行う。

2 関連研究

日本語固有表現認識の抽出手法の評価においては一般に IREX ワークショップ [6] による定義を用いる場合が多い。IREX では固有表現を PERSON、LO-

† 京都大学大学院情報学研究科

表 1: 固有表現のクラスとその例

クラス	例
PERSON	田中、木村庄之助
LOCATION	太平洋、東京都、中南米
ORGANIZATION	松下電器、自民党
ARTIFACT	ＰＬ法案、カローラ
DATE	２１世紀、昨年春
TIME	午前７時、正午
MONEY	５００億円、１２３カナダドル
PERCENT	２０%、５割

CATION、ORGANIZATION、ARTIFACT、DATE、TIME、MONEY、PERCENT の 8 クラスとして定義している。各クラスに属する固有表現の例を表 1 に示す。

日本語固有表現認識では主に、人手によって記述されたルールに基づく手法と、タグ付きコーパスを利用した機械学習に基づく手法の 2 つに分かれるが、先行研究では後者の機械学習を用いる手法の方が高精度を実現している。

一般に機械学習に基づく固有表現認識は、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。各固有表現に関して 1 形態素の固有表現 (S)、複数形態素の固有表現の先頭 (B)、途中 (I)、末尾 (E)、固有表現ではない (O) のいずれかのラベルを形態素列に対して付与する問題と考えることができる[1]。

このようなラベル付けは、コーパスに対して固有表現認識の正解ラベルを付与し、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。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。

系列ラベリングに基づく固有表現認識は局所的な情報のみを用いるため有用な情報を必ずしも全て利用することができない。そこで、大域的な情報を用いて固有表現認識を行う手法が提案されている。中野らは機械学習に用いる素性として文節内の解析方向に存在する固有名詞の品詞細分類に関する素性 (文節内素性)、解析方向に隣接する文節の末尾が名詞であった場合にその名詞 (隣接文節素性)、各文節の主辞に関する素性 (主辞素性) を用いている [4]。笹野らは構造的な素性としてキャッシュ素性や共参照解析、構文解析、格解析の結果を利用した素性を用いている [5]。

また、近年 Web などのタグ付けされていない大規模コーパスから知識を獲得し、それを固有表現認識に利用する研究が盛んに行われている。Kazama らは Wikipedia から抽出した固有表現辞書と web 文書から動詞と名詞の係り受け関係を収集しクラスタリングすることで作成した名詞クラスを利用している [9]。また福島らは、大規模な web テキストからパターンを利用して “政党-新党大地”,“企業-トヨタ” といった “カテゴリ名-固有名” ペアを獲得し利用している [10]。

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。

3 提案手法の概要

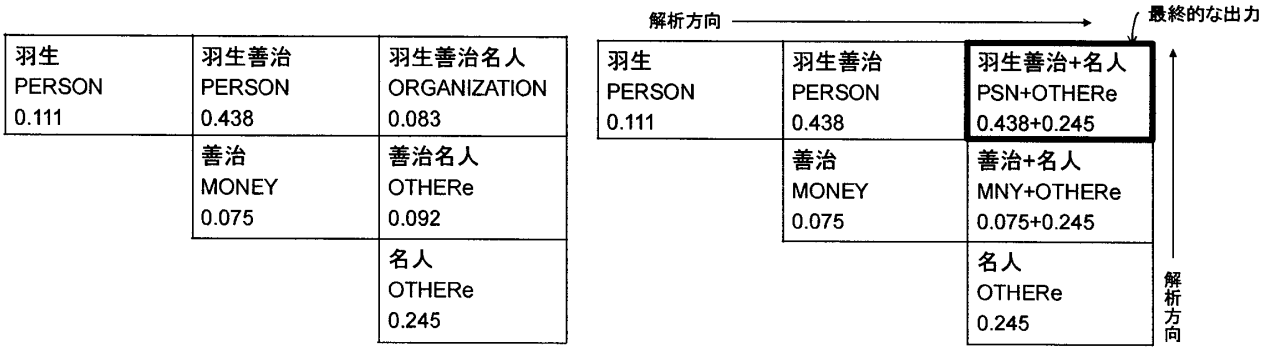
本手法では、ある範囲内であらゆる複合名詞のラベルを推定した後、CKY 法と同様の手法でボトムアップに解釈を決定することで固有表現認識を行う手法を提案する。「ある範囲」を以後、“解析単位” と呼ぶことにする。図 1 に提案手法の考え方を示す。例として “羽生善治名人” という文節を考え、この解析単位に対するラベルを決定する場合を考える。基本的な考え方は図 1(a) のように解析単位内の “羽生”, “羽生善治”, “羽生善治名人” など、1 つ以上の形態素から構成されるすべてのまとまり (チャンク) に対して機械学習によりラベルを推定する。

全てのチャンクのラベルを推定し、図 1(a) のように配置したものを初期状態と呼ぶことにし、次にボトムアップに解析単位内で最適なラベルを決定する (図 1(b))。この際、CKY 法のような考え方で図 1(a) の左下 (各形態素) から右上に向かって順にラベルを決定していく。例えば “羽生善治” という単位に対して “羽生善治” 全体を PERSON と解釈するのがよいか、“羽生” を PERSON、“善治” を MONEY と解釈するのがよいかというような比較を各セルについて行い、ラベル付けを決定していく。構文解析に用いられる一般的な CKY 法では文法規則を用いるのに対して、本手法は機械学習により各セルに対して最適なチャンクの組み合わせを選択する。

以上から学習すべきモデルとしては

- チャンクのラベルを推定するモデル (ラベル推定モデル)

[1] このようなラベル付けの手法を SE 法 [7] 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 [8] などがある。



(a):初期状態　　(b):最終的な出力

図 1: 提案手法の基本的な考え方 (例文節:羽生善治名人)

- 2 つのラベル付けを比較するモデル (ラベル付け比較モデル)

の 2 つである。この 2 つのモデル学習の詳細について以下で述べる。

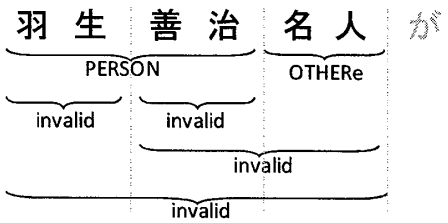


図 2: 文節 “羽生善治名人” に対するラベルの与え方

4 モデル学習

4.1 ラベル推定モデル

本節ではチャンクのラベルを推定するモデルを学習する手順を説明する。解析単位は基本的には先に述べたとおり文節とする。これはコーパスに出現する固有表現のうち 93.5%が 1 文節内にあるためである。ただし、例外的に “『ひめゆりの塔』”(ARTIFACT) のような括弧で括られた表現や “日本野鳥の会” (ORGANIZATION) のような Wikipedia にエントリがある表現はそれを解析単位として拡張する[2]。この拡張を行うことでコーパスに出現する固有表現のうち 98.6%が解析単位内に含まれる[3]。

このような解析単位から、各文節の先頭または末尾の機能語を削除する。例えば、“羽生善治名人は” という文節の場合、末尾の助詞 “は” を削除して “羽生善治名人” とし、“約三倍” という文節の場合、先頭の接頭辞を削除して “三倍” とし、この単位に対して学習を行う。

次に解析単位内のチャンクに対して以下のような形式でラベルを付与する。考えるラベルは IREX で定義された 8 種類の固有表現 (表 1 参照) に加えてそれ以外の表現に対する OTHERs、OTHERb、OTHERi、OTHERe、invalid の 5 種類の計 13 種類である。

OTHERs は解析単位全体がラベルなしの場合にその解析単位全体に与えるラベル、OTHERb、OTHERi、OTHERe はそれぞれ解析単位先頭、解析単位途中、解析単位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[4]。

IREX で定義された 8 種類の固有表現でもなく OTHER の 4 種類でもないチャンクに対しては invalid というラベルを与える。

例えば図 2 では “羽生善治” に PERSON,“名人” に OTHERe とラベル付けを行う。それ以外のチャンクには invalid を与える。

以上のようなラベルをあらゆるチャンクに付与することによって正解データを作成し、ラベルを推定するモデルを学習する。学習の素性には表 2 のものを用いる。(3),(5)−(8) に関しては「チャンク先頭」、「チャンク末尾」、「チャンク内のいずれかの位置に含まれる」の 3 つを区別する。例えば “羽生善治名人” というチャンクにおいて (7) の素性は先頭の形態素が “羽生”、末尾の形態素が “名人”、“善治” という形態素がチャンク内に含まれる、といったような素性を考える。また (11)−(14) 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する。

[2]wikipedia は 2007 年 10 月時点のデータを使用。

[3]1 つの解析単位内に含まれない例として “徹の君”(PERSON)、“大阪府緑の環境整備室”(ORGANIZATION) などがある。

[4]各 OTHER は形態素単位にラベル付けを行わず、解析単位内で最長の形態素列に対して与える。

表 2: ラベル推定の際に考える素性

(1) チャンクに含まれる形態素数
(2) 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
(3) 文字種[3]
(4) 隣接する形態素の文字種[3]の組み合わせ
  - 例えば “ロシア軍” というチャンクなら “片仮名, 漢字”
(5) JUMAN によって付与される品詞・品詞細分類・カテゴリなどの素性
(6) KNP によってチャンクを含む文節、形態素に付与される素性 [4]
(7) 文字列
  - チャンクに含まれる形態素、文節主辞
(8) IPADIC[5]に関する素性
  - チャンクが IPADIC の固有表現に関する分類 “人名”, “地域”,“組織”,“一般” のどれに属するか
(9) Wikipedia に関する素性
  - Wikipedia にエントリが存在するかどうか
  - 定義文から抽出した上位語 (例:“自民党” の上位語は “政党”)
(10) キャッシュ素性
  同一文書内で注目しているチャンクと同一のチャンクが以前に出現していた場合、以下のいずれか
  - 以前出現したチャンクのラベル
  - 以前出現したチャンクを内包するチャンクのラベル
(11) チャンクが属する文節に含まれる助詞
(12) 係り先の文節に含まれる形態素、助詞、文節主辞
(13) 用言格フレーム・名詞格フレームの固有表現・カテゴリに関する情報 [5]
  - 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
(14) 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
  - チャンクが括弧内にありかつその前に同格である表現がある場合、その同格表現の文字列

以上の各チャンクに対するラベル-素性から SVM を用いてモデルを学習する。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。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 個のモデルを学習する。SVM 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 13 種類のラベルのスコアの和が 1 になるように正規化する。

正例:
+1 羽生善治 vs 羽生 + 善治 ...(*)
PSN PSN + MNY
+1 羽生善治+名人 vs 羽生+善治, 名人
PSN+OTHERe PSN+MONEY,OTHERe
⋮
負例:
- 1 羽生善治名人 vs 羽生善治+名人
ORG PSN+OTHERe
⋮

図 3: 正例、負例の与え方

図 2 における “羽生” や “善治” というチャンクにおいては、invalid が比較的高いスコアを持つため PERSON の SVM のスコアが相対的に下がる。これにより、次節で述べるラベル付けの比較の際に “羽生善治” に PERSON とラベル付けする解釈が選ばれる。また invalid がスコア最大となったチャンクに対してはスコアが 2 位のラベルを採用する。これは、invalid とラベル付けされたチャンクを固有表現としてみなさない (全て OTHER として扱う) ようにすると、予備実験において再現率が低下し結果 F 値が低下したためである。

4.2 ラベル付け比較モデル

本節ではある範囲での 2 種類の解釈 (ラベル付け) のいずれがよいかを比較するモデルを学習する。例えば “羽生善治” という単位に対して “羽生善治” 全体を PERSON とする解釈と、“羽生” を PERSON、“善治” を MONEY とする解釈を比較し、前者を選択する。

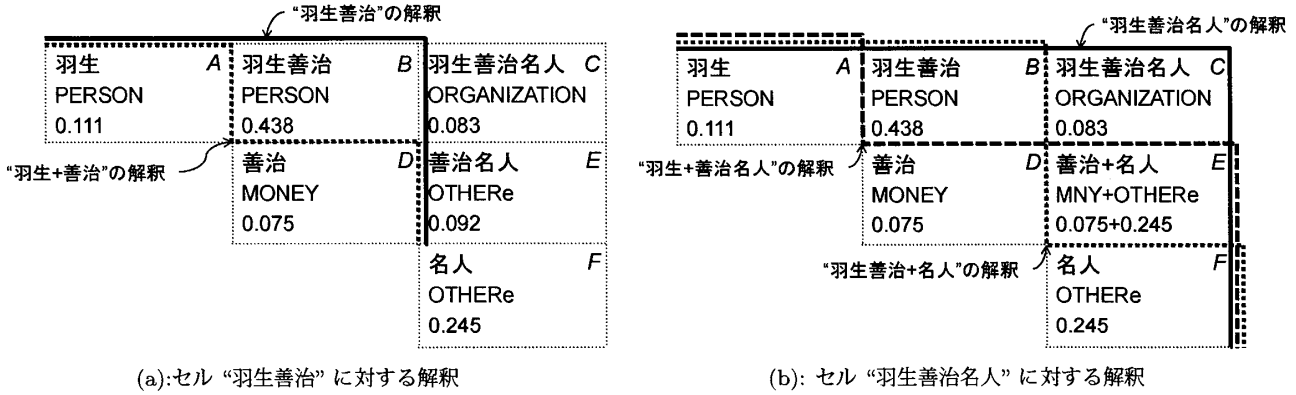
まず、図 3 のように比較したいチャンクの組み合わせを「vs」をはさんで 2 つ並べる (左から 1 個目、2 個目と呼ぶことにする)。1 個目のチャンクの組み合わせが正解であれば正例、2 個目のチャンクの組み合わせが正解であれば負例とする。また 1 個目、2 個目ともにチャンク数の最大許容数はそれぞれ 5 とし、チャンク数が 6 以上となるものは考慮しないことにした。

次に、各サンプルに対する素性の与え方について図 3 の (*) を例にとり説明する。各チャンクに対する素性ベクトルを 13 次元とする。この素性ベクトルは以下のように定める。まず 12 個のラベルと次元との対応を決めておき、対応するラベルの次元の素性に 1 を

[3]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。
[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。
[5]ipadic version 2.7.0(http://sourceforge.jp/projects/ipadic/)

チャンク	羽生善治					羽生	善治			
ラベル	PERSON					PERSON	MONEY			
ベクトル	$\mathbf{V_{11}}$	**0**	**0**	**0**	**0**	$\mathbf{V_{21}}$	$\mathbf{V_{22}}$	**0**	**0**	**0**

図 4: “羽生善治 vs 羽生+善治” に対する次元数の確保の仕方 ($\mathbf{V_{11}}$, $\mathbf{V_{21}}$, $\mathbf{V_{22}}$, **0** はそれぞれ 13 次元のベクトル)



(a):セル “羽生善治” に対する解釈　　(b): セル “羽生善治名人” に対する解釈

図 5: ラベルを決定する際の比較

立てる。また、13 次元目は対応するチャンクのラベル付けをした際の SVM の出力スコア (ラベル推定モデルのスコア) を素性として用いる。そして、1 個目、2 個目のチャンクの組み合わせそれぞれにおいて順に左からチャンクの素性ベクトルを並べ、チャンクの数が最大許容数 5 未満である場合は余ったところに零ベクトルを置く。

図 4 に “羽生善治 vs 羽生+善治” に対する素性ベクトルの次元の確保の仕方を示す。この例の 1 個目のチャンクの組み合わせは、「羽生善治」という 1 つのチャンクから構成されるため、先頭の素性ベクトルはこれに対応するベクトルとし、それ以降の 4 つの素性ベクトルは対応するチャンクがないため零ベクトルとする。このようにして作成した正例・負例と素性のペアから SVM を用いてモデルを作成する。

なおこの例の場合、正解データからは “羽生善治” が PERSON であることしかわからないので例えば “羽生善治名人 vs 羽生+善治名人” の場合はどちらが正解かを判断できない。したがってモデルの学習に用いることができない。

5 解析

解析は、まず解析単位内のあらゆるチャンクに対して 4.1 節で作成したラベル推定モデルを用いてラベルを決定する。次に 4.2 節で作成したラベル付け比較モデルを用いて解析単位内で最適なラベルの組み合わせを決定する (図 1)。その際、構文解析に用いられる CKY 法のようにボトムアップに解析単位のラベルを決定していく。

ラベル推定モデルによって図 1(a) が得られたとする。そこからラベル付け比較モデルを用いて表の対角線上、すなわち形態素単位のセルから順にラベルを決定していく。“羽生善治” のセルに対する最適なラベル付けを決定する際には、図 5(a) のように B と解釈するのがよいか A+D と解釈するのがよいのかを比較する。この場合 B の単位すなわち “羽生善治” を PERSON とするラベル付けが採用される。同様に “善治名人” のセルに対する最適なラベル付けを決定する場合も E と解釈するのがよいか D+F と解釈するのがよいのかを比較し、この場合 D+F すなわち “善治” を MONEY、“名人” を OTHERe とするラベル付けが採用される。より短い単位からなる組み合わせが採用されたらセル (この場合図 5(b) の E の部分) の内容を更新する。

また最も右上の部分、つまり最終的なラベル付けを決定する際には、図 5(b) のように A+E(この場合 D+F) という解釈、B(この場合 B)+F という解釈、C という解釈の 3 つの組み合わせを考える必要がある。解釈が 3 つ以上ある場合はすべての解釈の組み合わせについて pairwise に比較を行い、ラベル付け比較モデルのスコアの和が最も高いものを採用する。

このラベル付け比較の際には、OTHER - OTHER の隣接をゆるさないという制約条件を与えている。これは OTHER は文節内で最長になるようにモデルを考えたためである。また以下のように 1 個目と 2 個目の解釈が等しくなるような場合は比較を行わない。

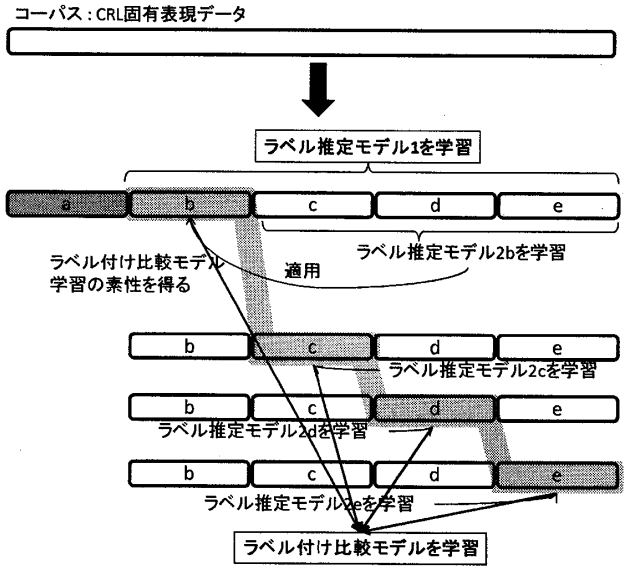


図 6: 5 分割交差検定の方法

新民主+ 連合所属, 議員　vs 新民主, 連合所属+議員
ART + LOC,OTHERe vs ART, LOC + OTHERe

以上のような手順で最終的なラベル付けが決定される。後処理として例えば “三郷/市 長” のように形態素より短い固有表現を先行研究 [5, 11] と同様にルールベースで分割する。これにより地名 “三郷市” が抽出できる。ルールは他に “半年”, “半数” を “半” (PERCENT)、“改憲”, “違憲” を “憲” (ARTIFACT) とするものなどがある。

6 実験

CRL 固有表現データを用いて実験を行った。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、先行研究ではこのような表現に対する評価は行っていないため、本研究においてもこのような表現は学習には用いず[6]、評価する際も対象外とした。

本実験では従来研究と同様に 5 分割交差検定を行った。SVM によるモデルの学習を 2 度行う必要があるため、図 6 のようにコーパスを分割する。5 分割した a の部分の固有表現認識を行う場合を考える。まず b-e の部分でラベル推定モデル 1 を学習する。次に b-e を 4 分割する。c-e の部分のみでラベル推定モデル 2b を学習しそれを b の部分に適用することでラベル付け比較モデル学習ための素性を得る。同様に b,d,e からラベル推定モデル 2c、b,c,e からラベル推定モデル 2d、b-d からラベル推定モデル 2e を作成し、それぞれ c,d,e の部分に適用してラベル付け比較モデル学習のための素性を得て、全体を用いてラベル付け比較モデルを学習する。そしてラベル推定モデル 1 とラベル付け比較モデルを用いて a の部分の解析を行う。

[6]ただしラベル推定モデル学習する際、4 種類の OTHER や invalid の学習には用いた。

表 3: 固有表現認識の結果

	Recall	Precision
ORGANIZATION	81.39 (2992/3676)	87.87 (2992/3405)
PERSON	89.27 (3428/3840)	93.87 (3428/3661)
LOCATION	91.52 (5000/5463)	92.68 (5000/5395)
ARTIFACT	47.26 (353/ 747)	74.16 (353/ 476)
DATE	93.58 (3338/3567)	93.61 (3338/3566)
TIME	88.84 (446/ 502)	90.47 (446/ 493)
MONEY	93.85 (366/ 390)	94.57 (366/ 387)
PERCENT	94.51 (465/ 492)	94.70 (465/ 491)
ALL-SLOT	87.74	91.69
F-measure		89.67

本実験を行うにあたり、形態素解析器として JUMAN[7]、構文解析器として KNP[8]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 3 に示す。全体の精度は F 値で 89.67 であった。先行研究の中で最も精度の高い笹野らの手法 [5] となるべく実験条件を揃えるために Wikipedia を利用しなかった場合の精度を算出したところ、F 値は 89.53 であった。

また、ラベル推定モデルのみで推定したあらゆるチャンクに対する Recall は 90.30 であった。つまり、この Recall と 100 との差 (9.70) がラベル推定モデルの誤りを表し、表 3 における ALL-SLOT の Recall との差 (2.56) がラベル付け比較モデルの誤りを表す。

7 考察

7.1 先行研究との比較

本手法では、笹野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

[7]JUMAN version 6.0(http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html)
[8]KNP version 3.0(http://nlp.kuee.kyoto-u.ac.jp/nl-resource/knp.html)

表 4: 先行研究との比較

	CRL 交差検定	解析単位	その他の素性
磯崎ら 2003[11]	86.77	形態素	
浅原ら 2003[12]	87.21	文字	
中野ら 2004[4]	89.03	文字	文節素性
山田 2007[13]	88.33	形態素	文節素性
笹野ら 2008[5]	89.40	形態素	構造的情報
風間ら 2008[9]	88.93	文字	Wikipedia, ウェブテキスト
福島ら 2008[10]	89.29	文字	ウェブテキスト
提案手法	**89.67**	複合名詞	Wikipedia, 構造的情報
提案手法 (-wikipedia)	89.53	複合名詞	構造的情報

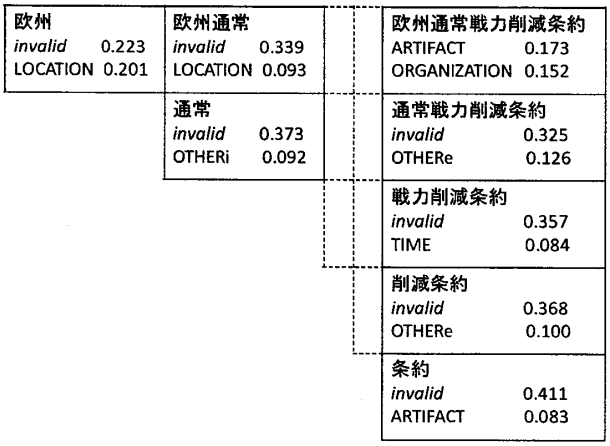


図 7: “欧州通常戦力削減条約” に対するラベル (一部)

笹野らの手法では “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。このときのラベル付けの一部を図 7 に示す。図 7 では、各セルにスコアが 1 位、2 位のラベルを invalid も含めて表示する。笹野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おり、この文節の主辞は “条約” であるので “欧州” のラベルを推定する際に “条約” から得られる情報を用いている。しかし “欧州” を LOCATION とする判断を覆すのに十分でなかったと推測できる。これに対して図 7 を見ると本手法では、“欧州” のラベルとして 1 位が invalid となり LOCATION のラベルのスコアを十分に下げ、最終的な出力も正解になったと考えられる。また、“戦力削減条約”’ というチャンクに対しても invalid が高いスコアになっており、“欧州戦力削減条約” という単位のスコアが相対的に向上している。

また “外国人登録法違反の” という文節において笹野らの手法では “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。この例では文節の主辞が “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。

先行研究との比較を表 4 に示す。本手法は、既存研究よりも高い精度であることがわかる。先行研究の中で高い精度が報告されているもののうち、福島ら [10] の手法は、大規模な web テキストからパターンを利用して “政党-新党大地”, “企業-トヨタ” といった “カテゴリ名-固有名” ペアを獲得し、利用している。また、笹野らの手法では本手法では用いていない共参照解析の結果を素性として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。

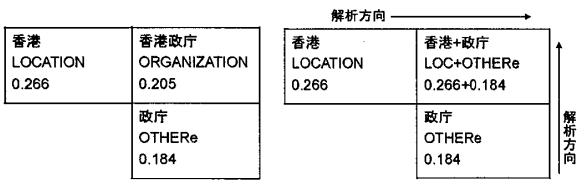
7.2 エラー分析

まずカタカナ語に関する誤りが多く見られた。例えば以下の例 (4) においては “バティストゥータ” を PERSON と認識するのが正解であるが、システムは OTHERs と出力する。

(4) イタリアで活躍するバティストゥータを加えたアルゼンチン。

“バティストゥータ” に関しては JUMAN の辞書や Wikipedia にエントリがなく、前後の係り受けや格解析の結果などの構文的な情報から判断するしかない。しかし格解析において “活躍する” のガ格を決定する際に、“バティストゥータ” が未知語であるため解析誤りが生じ、それが原因で固有表現認識誤りが生じたと考えられる。

また本稿ではチャンクの解析は正しいが、解析単位内での最も良い解釈を選択する際に誤ったものが存在



(a):初期状態　　(b):最終的な出力

図 8: エラー分析の例

した。例えば “香港政庁” という文節において “香港政庁” に ORGANIZATION とラベル付けされるのが正解であるが、図 8(b) のように “香港” に LOCATION とラベル付けされる。同図 (a) において “香港政庁” 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、“香港政庁” と “香港+政庁” で “香港+政庁” の方が良いと判断されたことがわかる。

このような誤ったラベル付けを減らす方法としてシグモイド関数の β の値の調整、2 つのモデルに対する素性の工夫などが考えられる。

8　まとめと今後の課題

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、さらに推定されたラベルからボトムアップの手法を用いて最適なラベルを SVM を用いて決定することにより固有表現認識を行った。CRL 固有表現データを用いて 5 分割交差検定を行ったところ F 値 89.67 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 [14] へ組み込む予定である。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。したがって個々のチャンクから最終的な出力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。構文解析と固有表現解析を統合させる研究としては、Finkel[15] らが英語の構文木に固有表現のタグを付与し、それを素性として用いることで双方の解析精度が向上すると報告している。

また、SVM は分類処理の速度が遅いという問題があり、提案手法では Xeon 2.33GHz の CPU を用いて解析時間が 0.973 文/秒であった。これに対して SVM を用いた山田 [13] の手法では PentiumIII 833MHz の CPU で 0.8 文/秒、磯崎ら [11] の手法では Atholon 1.3GHz の CPU で 4,806 バイト/秒であり、高速化は今後の課題である。

参考文献

[1] Vladimir Vapnik, *The Nature of Statistical Learning Theory*, (Springer, 1995).

[2] John Lafferty, Andrew McCallum, and Fernando Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data, *Proceedings of ICML'01*, (2001), pp. 282–289.

[3] Vajay Krishnan and Christopher D.Manning, An effective two-stage model for exploitiong non-local dependencies in named entity recognition, *Proceedings of COLING-ACL'06*, (2006), pp. 1121–1128.

[4] 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.

[5] Ryohei Sasano and Sadao Kurohashi, Japanese named entity recognition using structural natural language processing, *Proceedings of IJCNLP'08*, (2008), pp. 607–612.

[6] IREX Committee (editor), *Proceedings of the IREX Workshop*, (1999).

[7] Satoshi Sekine, Ralph Grishman, and Hiroyuki Shinnou, A decision tree method for finding and classifying names in Japanese texts, *6th Workshop on Very Large Corpora (WVLC-6)*, (1998), pp. 171–178.

[8] Erik F. Tjong Kim and Jorn Veenstra, Representing text chunks, *Proceedings of EACL'99*, (1999), pp. 173–179.

[9] Jun'ichi Kazama and Kentaro Torisawa, Inducing gazetteers for named entity recognition by large-scale clustering of dependency relations, *Proceedings of ACL-08: HLT*, (2008), pp. 407–415.

[10] 福島健一, 鍛治伸裕, 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, *DEWS 08*, (2008), A3-3.

[11] 磯崎秀樹, 賀沢秀人, 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.

[12] Asahara Masayuki and Yuji Matsumoto, Japanese named entity extraction with redundant morphological analysis, *Proceedings of HLT-NAACL'03*, (2003), pp. 8–15.

[13] 山田寛康, Shift-reduce 法に基づく日本語固有表現抽出, 情報処理学会自然言語処理研究会 2007-NL-179-3, (2007), pp. 13–18.

[14] Daisuke Kawahara and Sadao Kurohashi, Probabilistic coordination disambiguation in a fully-lexicalized japanese parser, *Proceedings of EMNLP-CoNLL'07*, (2007), pp. 304–311.

[15] Jenny Rose Finkel and Christopher D.Manning, Joint parsing and named entity recognition, *Proceedings of NAACL/HLT'09*, (2009), pp. 326–334.