品詞タグ付きコーパスの統計情報と構文・格・文脈情報を統合した固有表現認識

河原 大輔　　黒橋 禎夫

京都大学大学院情報学研究科

Named Entity Extraction based on Statistical Information in a Tagged Corpus and Syntactical, Subcategorizational, and Contextual Information

Daisuke Kawahara　　Sadao Kurohashi

Graduate School of Informatics, Kyoto University

Abstract: Named entities cannot be extracted accurately even if a wide-coverage dictionary is prepared. This is because unknown words appear very often, and there are many ambiguities among person, location and organization names. The human, however, can recognize named entities with no difficulty, using several types of cues in sentences. This paper proposes an integrated method to extract named entities based on such cues as named entity probability of a word, kana-kanji types of the word, phrases before and after the word, phrasal patterns of named entities, and syntactical, subcategorizational, and contextual information.

1　はじめに

固有表現の処理は、情報検索・抽出などにおいて重要であるだけでなく、形態素・構文解析を高度化するという言語処理の基礎的課題においても重要である。すなわち、一般的な意味で文法を整備しただけのシステムでは固有表現の扱いが十分に行えず、それが形態素・構文解析の中の一つの障害となるからである。

固有表現は、ある国の見知らぬ町の名前や、新たに発生する人名など、基本的に未知語に対する処理を必要とする。さらに、地名と人名、また組織名の間で多くの曖昧性がある。これらのことから、辞書を整備するという方法だけでは固有表現をうまく処理することはできない。

これに対して、文章中には固有表現に関するさまざまな手がかりがあり、人間の場合にはそれらをみることでほとんどの場合正しく固有表現を認識することができる。本稿では、人間が行っているように、さまざまな手がかりを統合することによって、すなわち、単語の固有表現の可能性、文字種、前後の表現、句としての定型的パターン、さらに構文、格、文脈情報などを用いて固有表現を認識する方法を提案する [6]。

2　枠組み

我々の手法では、まず、形態素単位の固有表現のタグ付けを行い、次にその結果を組み合わせて形態素列(句) 単位の固有表現タグ付けを行う。例えば、「京都市左京区で…」という表現では、まず「京都」と「左京」に形態素としての地名のタグを与え、さらにそれらのまとまりの「京都市左京区」に句としての地名のタグ

を与える。以下では、前者の処理を「形態素固有タグ付け」、後者の処理を「句固有タグ付け」と呼ぶことにする (IREX–NE の課題はこの句固有タグ付けにあたる)。

形態素固有タグは、人名 (PERSON)、地名 (LOCATION)、組織名 (ORGANIZATION)、固有物名 (ARTIFACT)、その他 (OTHER) の 5 つであり、句固有タグは、その 5 つに日付表現 (DATE)、時間表現 (TIME)、金額表現 (MONEY)、割合表現 (PERCENT) を加えたものである。

ひとつの記事に対するタグ付けの手順は以下のとおりである。

1. 一文の形態素解析と構文解析を行う (JUMAN[4], KNP[3])。
2. 一文の形態素固有タグ付けを行う。
   (a) 形態素の前後の情報を統合して、決定木によるタグ付けを行う。
   (b) 記事中の前文までのタグ付けの結果 (文脈情報) を用いてタグ付けを行う。
3. 一文の句固有タグ付けを行う。
   (a) 2 の結果を用いてルールによるタグ付けを行う。
   (b) 3a でタグ付けされていないものについて、構文・格情報を用いたタグ付けを行う。
4. 上記 1、2、3 の処理を記事の終わりまで繰り返す。

3 形態素固有タグ付け

3.1 決定木によるタグ付け

固有表現認識の第一段階として、文中の各形態素に対して固有タグを付ける。ある形態素が固有表現であるかどうかを判定するための情報は、その形態素自身、あるいは、その直前、直後にあることが多い。例えば、多くの場合未知のカタカナ語は固有表現である。また、「山口市」であれば「山口」は地名、「新人の山口」であれば「山口」は人名であるとわかる。

ここで問題となるのは、これらの情報をどのように組み合わせて利用するかということで、これを人手でルールとして記述するのは困難である。そこで、コーパス中の形態素に対して正解の固有表現タグを与え、情報の取捨選択方法を決定木によって学習することとした。各形態素に与える素性は以下のものとした。

対象形態素の出現頻度

学習コーパス中での、その形態素が出現した回数

対象形態素の文字種

かな漢字 (ひらがな、漢字、または漢字とひらがなの混合)、カタカナ、英記号 (アルファベット、または記号)、数字のいずれか

対象形態素のタグ分布

学習コーパス中において、その形態素がどのようなタグを持っているかの百分率

例: 「山口」 ··· 人名:81%, 地名:18%, 組織名: 0%, 固有物名:0%, その他:0%

対象形態素の形態素解析結果におけるタグリスト

形態素解析 (JUMAN) の結果として与えられるタグの可能性

例: 「山口」 ··· 人名:有, 地名:有, 組織名:無, 固有物名:無, その他:無

隣接形態素によるタグ分布

隣接または、接続助詞 “の” でつながっている前後の形態素 (名詞または接辞) によって、対象形態素 (文字種で区別) のタグがどのような分布になるか

例: X(かな漢字)「氏」 ··· 人名:98%, 地名:0%, 組織名: 0%, 固有物名:0%, その他:1%

例: X(かな漢字) の「大敗」 ··· 人名:0%, 地名:0%, 組織名: 75%, 固有物名:0%, その他:25%

例: 「東南」X(カタカナ) ··· 人名:0%, 地名:95%, 組織名: 0%, 固有物名:0%, その他:4%

例: 「評論家」の X(かな漢字) ··· 人名:87%, 地名:0%, 組織名: 0%, 固有物名:0%, その他:12%

学習データとしては京都大学テキストコーパス [5] 約 25,000 文を用い、決定木の構築には C4.5[1] を用いた[1]。

3.2 文脈情報の利用

新聞記事などでは、固有表現が最初に導入される場合には省略のない表現が用いられ、二回目以降は省略表現が用いられることが多い。この場合、二回目以降の省略表現の認識は難しい。例えば、日本人の人名は、次の例のように最初に姓名が述べられ、二回目からは姓だけで呼ばれることが多い。

(1) その中で、名人位に六度挑戦していずれも敗れていた米長邦雄が七度目の挑戦にして悲願の名人になった。… 昨日、米長に挑んだのが羽生。

[1] 実験では、信頼度 5% の枝刈りを行った (2%、5%、25% の中で 5% のときもっとも精度がよかった)。

このような省略表現をうまく認識するために、記事中で前に出現した固有表現との間で前方一致を調べ、一致するものがあれば、その形態素固有タグを用いる。ただし、カタカナ・カタカナという形の人名の場合は、後ろの語だけの一致も調べる (この結果は決定木による結果よりも優先する)。

4 句固有タグ付け

4.1 ルールによるタグ付け

句の固有表現には定型的なものが少なくない。そこで、それらのパターンをルールとして記述することによって句の固有タグ付けを行った。用いたルールの例を表 1 に示す。

このルールの条件部では、3 章の処理によって与えられた形態素固有タグに加えて、NTT 日本語語彙大系 [2] の意味属性の情報を用いる。例えば表 1 のルール 1 は、形態素固有タグが地名である語と意味素性が《地名末尾》[2]である語が連続すれば、それらに句固有タグとして地名を与えるということを意味する (ルールの条件部の下線の形態素列に、矢印の右側のタグを与える)。

句のまとまりのパターンに対してルールを考えることによって、その中の各形態素の認識が不完全であっても、全体のタグを推測することができる。例えば、「アントノフ・ソ連軍参謀総長」の場合、「アントノフ」の形態素固有タグが正しく求められていなくても、「アントノフ」がカタカナ、「参謀総長」が《役職》であるという情報から、「アントノフ」を人名、「ソ連軍」を組織名と認識することができる (表 1 のルール 6)。

4.2 構文・格情報の利用

前節のルールによるタグ付けは句の内部の手がかりに基づく処理であった。しかし、句の内部を見るだけでは十分な手がかりがない場合がある。そのような場合には、構文解析の結果を利用することによって、句の外側の情報に基づくタグ付けを行う。

構文解析の結果が利用できる場合の一つとして、並列構造の情報がある。つまり、並列関係にある文節間において、いずれかの文節が固有表現であれば、他方の文節に同じタグを与える。

(2) 香港の キャセイ航空(組織名) や台湾の 中華航空(?)

この例では、“キャセイ航空” が組織名とわかるので、それらと並列関係にある “中華航空” のタグとして組織名を与える。

この処理は、並列構造の解析が誤っている場合には副作用をおこす。そこで、並列構造の解析の信頼度が高い場合、すなわち、並列が 3 つ以上の句からなる場合 (上記の例のような場合)、あるいは 2 つの句の並列で、それぞれの句の長さが 2 文節以上の場合に限り適用することとした。

構文解析の結果が利用できるもう一つの場合として、構文的つながりが明らかになった文節間の意味的制約がある。例えば、動詞とその格要素の間には選択制限の制約がある。また判定詞およびで同格表現でつながる名詞同士は同じ意味タイプをもつと考えられる。

ただし、現在のところこのような意味的制約を高い信頼度で適用することは難しいので、これらは以下のように制限して適用することとした。

1. タグを与える対象の名詞句は、カタカナ・カタカナ、または (JUMAN の辞書において) 普通名詞の可能性のない漢字一単語とする。
2. 動詞の選択制限は、NTT 日本語語彙大系の構文体系において、ガ格またはヲ格が、《人名》または《組織名》のみである場合に、そこに人名または組織名のタグを与える。
3. 判定詞の構文「Y が‥ X だ」、同格表現「X の Y」においては、X が《人名》または《組織名》の意味属性をもつ場合、Y に人名または組織名のタグを与える。

以下にこれらのタグ付けが適用される例をあげる。

(3) a. リロイ・バレル (人名) が…更新した。
b. ジャンフランコ・マシャ(人名) 事務局長は…
ごく普通の市民。
c. 新横綱の 貴乃花(人名) は…

5 結果と考察

IREX–NE の本試験の結果は F-Measure 値で、総合ドメイン 71.96、逮捕ドメイン 72.77 であった。本試験時のシステムには、人名の読みなどの括弧の扱いの不備、住所の認識ルールの不備、若干のバグなどがあった。それらの修正を行った結果、総合ドメイン 75.25、逮捕ドメイン 78.91 となった。

[2]《地名末尾》は実際には《行政区画》、《領土》、《公共施設》、《宗教施設》、《居住施設》の集合である。

表 1: ルールの例

	ルール	例
1	[地名] 《地名末尾》 → 地名	大阪湾
2	[地名 | 組織名 | 人名] 《組織名末尾》 → 組織名	ドゥダエフ政権部隊
3	《路線名》 ? “駅” → 地名	ＪＲ東京駅
4	カタカナ・《独立国名》 《役職》 → 人名	エリツィン・ロシア大統領
5	漢字$_1$・?$_2$ 《役職》 → 1:人名, 2:組織名	黒木三郎$_1$・早大$_2$ 名誉教授
6	カタカナ$_1$・?$_2$ 《役職》→ 1:人名, 2:組織名	アントノフ$_1$・ソ連軍$_2$ 参謀総長
7	?$_1$ [組織名 | 地名]$_2$ 《役職》→ 1:人名, 2:組織名	松下康雄$_1$日銀$_2$ 総裁

? はひとつ以上の形態素を表す。

5.1 認識誤りの傾向

本手法の認識誤りの傾向を以下に挙げる。

- 句固有タグ付けのルールの柔軟性の問題

  例: 国際航空運送協会(組織名) (ＩＡＴＡ(組織名))

  “国際航空運送協会” には “協会” しか手がかりがないため現在のルールでは認識されないが、例えばある長さ以上の複合語であるとか、後ろに括弧付きのアルファベット列があるなどの条件によってより柔軟に組織名を推測することが考えられる。

- 文脈情報の柔軟性の問題

  例: 長銀(組織名) をめぐっては、不良債権を過小評価して…

  “長銀” は同記事中に出てきている “日本長期信用銀行” の略語であるが、このような略語は現在の文脈情報の処理のマッチングでは扱うことができない。

- 固有物名 (ARTIFACT) の問題

  例: シャガールの「花を囲んで踊る人(固有名詞)」

  固有物名の認識はもっとも難しい。固有物名は引用符に囲まれていても用言を含む場合は、単純に固有物名とすると誤認識する可能性が大きい。従って、より高次の情報を利用することが必要である。

5.2 構文・格・文脈情報の利用による効果

文脈情報、並列構造、構文による意味制約によって付けられたタグの数はそれぞれ、289 個、28 個、9 個であった。全体の正解タグ数は 1897 個であったので、並列構造、構文による意味制約によるタグ付けの適用回数は非常に少ないという結果であった。

これは、信頼性の高い場合にのみこれらの情報を用いるというように適用条件を非常に厳しくしたためである。構文解析の精度の向上、格フレーム辞書の整備などを進めることによってこれらの高次の処理の信頼性を高めることが必要である。現在のシステムでは、形態素・構文解析と固有名詞認識処理をそれぞれ別々に行っているが、これらを統合することによってそれらの精度を相補的に高めることが今後の課題である。

参考文献

[1] J.Ross Quinlan. AI によるデータ解析. トッパン, 1995.

[2] NTT コミュニケーション科学研究所. 日本語語彙大系. 岩波書店, 1997.

[3] 黒橋禎夫. 日本語構文解析システム KNP version 2.0 b6 使用説明書. 京都大学大学院 情報学研究科, 6 1998.

[4] 黒橋禎夫, 長尾真. 日本語形態素解析システム JUMAN version 3.6 使用説明書. 京都大学大学院 情報学研究科, 11 1998.

[5] 黒橋禎夫, 長尾眞. 京都大学テキストコーパス・プロジェクト. 言語処理学会 第 3 回年次大会発表論文集, pp. 115–118, 1997.

[6] 大石巧, 黒橋禎夫, 長尾眞. コーパス中の特徴と文法的意味的情報を統合的に用いた新聞記事中の固有名詞認識. 言語処理学会 第 4 回年次大会発表論文集, pp. 43–46, 1998.