TR－SLT－0050

ウェブ上のテーブルからの固有表現抽出
Named Entity Extraction from Tables on The Web

佐々木 靖弘　　菊井 玄一郎
Yasuhiro Sasaki　　Genichiro Kikui

２００３年９月１９日

概要

音声認識のための語彙を獲得する方法として、本研究ではウェブ上のテーブルから固有表現を抽出する方法を提案する。ウェブのテーブルでは、同じジャンルの固有表現が同じ列に現れるケースが多く見られる。提案手法では、シードワードとなる固有表現をシステムに与えて、シードワードと同じテーブルで同じ列に現れる表現を固有表現として抽出する。また、抽出される固有表現は頻度情報に基づく確信度によって順位付けられており、確信度上位の表現をシステムの新たな入力とすることにより、繰り返し固有表現を抽出することが可能である。

（株）国際電気通信基礎技術研究所
音声言語コミュニケーション研究所
〒619－0288「けいはんな学研都市」光台二丁目２番地２ TEL：0774－95－1301

Advanced Telecommunication Research Institute International
Spoken Language Translation Research Laboratories
2-2-2 Hikaridai "Keihanna Science City" 619-0288,Japan
Telephone:+81-774-95-1301
Fax　　:+81-774-95-1308

目次

第1章

はじめに

実際の会話では、映画名や料理名などの分野特有の固有名詞、あるいは、固有名詞的機能を持つ連語が頻繁に現れる。これらの固有表現は情報を伝達する上で重要な意味を持つため、音声言語処理システムを構築する際には、対象分野の固有表現に関する知識が必要である。たとえば、通常の音声認識処理においては認識用の辞書に対象単語が全て含まれていなければならない。また、未登録語（未知語）を統計的にモデル化する手法においても、学習データとして当該分野の単語のリストが必要である。

ところが、氏名や地名など一部のカテゴリを除いて、固有表現を網羅的に収集したものは少ない。また、固有表現は日々新語が産み出されるため、コストをかけず継続的に収集する必要がある。そこで、本研究では、ウェブ上のテーブルから固有表現を自動的に抽出する方法を提案する。

ウェブには数多くの分野の固有表現が存在し、なおかつ、新しく生まれた表現がウェブには比較的早く出現するということから、これらを効率よく収集することが期待できる。

ウェブ上のテーブルにおいて、ある固有表現が現れる列には、その固有表現と同じジャンルの固有表現が現れることが期待される。本研究で作成する固有表現抽出システムでは、入力として数語のシードワードとなる固有表現を与えると、ウェブ上のシードワードを含むテーブルから、シードワードと同じ列に現れる表現を固有表現として抽出する。

また、抽出される固有表現は頻度情報に基づく確信度によって順序付けられており、確信度上位の固有表現をシステムの新たな入力とすることにより、繰り返し固有表現を抽出することが可能である。

以下では、2章で本研究で作成した固有表現抽出システムについて説明する。3章では、作成した固有表現抽出システムで行った固有表現の抽出実験について述べる。最後に、4章でまとめと今後の課題について述べる。

第 2 章

固有表現抽出システム

本研究で作成した固有表現抽出システムは、以下の 3 つの部分から構成される。

1. HTML ファイル収集部
2. 固有表現抽出部
3. 確信度付与部

図 2.1 に、システムの構成図を示す。

入力として、抽出したいジャンルに含まれる固有表現をシードワードとして数語与えると、システムはシードワードと同ジャンルの固有表現を、確信度という数値で順序付けて出力する。固有表現を順序付けて出力することができれば、確信度上位の数語をシステムの新たな入力とすることにより、何度も固有表現の抽出を繰り返すことが可能となる。

本章ではまずこれら 3 つの部分について述べ、次に固有表現抽出の繰り返しについて述べる。

2.1 HTML ファイル収集部

本システムでは、抽出したいジャンルに含まれる数語の固有表現をシードワードとして入力する。HTML ファイル収集部では、入力されたシードワードをサーチエンジンのクエリとして AND 検索を行い、得られた結果の上位 n 位の URL から HTML ファイルをダウンロードする。サーチエンジンには goo [*1] を利用している。

このとき、ジャンルを限定するための語をクエリとして AND 検索に加えることができる。例えば、映画のタイトルを抽出する際に、「映画」という言葉を AND 検索に加えることができる。

[*1] http://www.goo.ne.jp/

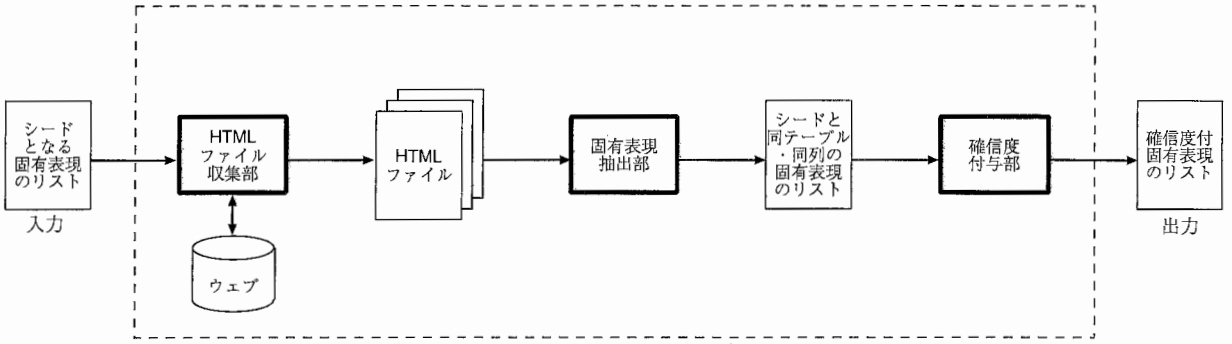


図 2.1: 固有表現抽出システムの構成図

2.2 固有表現抽出部

固有表現抽出部では、HTML ファイル収集部で収集した HTML ファイルを解析し、シードワードと同じテーブルで同じ列に現れる表現を固有表現として抽出する。ここでは、以下の 3 段階の処理を行っている。

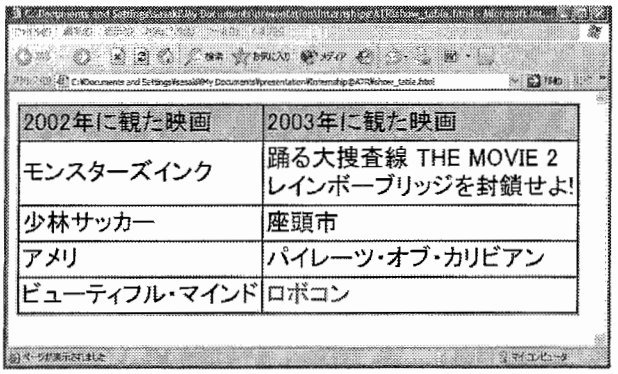
1. HTML ファイルからのテーブルの抽出
2. コンテンツの抽出と番号の付与
3. シードワードと同列の固有表現の抽出

2.2.1 HTML ファイルからのテーブルの抽出

固有表現抽出部ではまず、ダウンロードした HTML ファイルからテーブルのみを抽出する。HTML でテーブルを記述する際には、主に <table> タグ、<tr> タグ、および <td> タグが用いられる。<table> タグはテーブル全体を記述するのに用いられ、<tr> タグは 1 つの行を記述するのに用いられ、<td> タグは 1 つのセルを記述するのに用いられている。この他に見出しを記述するための <th> タグや、テーブルをヘッダとボディとフッタに分けるためのタグなどが存在するが、本システムでは利用しない。

例えば、ウェブブラウザで図 2.2 のように表示されるテーブルの HTML ファイルは図 2.3 のように記述される。

この HTML ファイルから <table>、</table> タグに囲まれた、<table> タグ、<tr> タグ、<td> タグおよびテーブルのコンテンツを抽出する。コンテンツとは、ここではセルの内容を意味する。図 2.2 のテーブルのコンテンツは『2002 年に観た映画』、『2003 年に観た映画』、『モンスターズインク』などである。



2002年に観た映画	2003年に観た映画
モンスターズインク	踊る大捜査線 THE MOVIE 2 レインボーブリッジを封鎖せよ!
少林サッカー	座頭市
アメリ	パイレーツ・オブ・カリビアン
ビューティフル・マインド	ロボコン

図 2.2: テーブル例

```
<html>
<body>
<table cellspacing='1' cellpadding='3' border='1'>
<tr bgcolor='#D3D3D3'>
<td>2002年に観た映画</td><td>2003年に観た映画</td>
</tr>
<tr>
<td>モンスターズインク</td>
<td>踊る大捜査線 THE MOVIE 2<br>
レインボーブリッジを封鎖せよ!</td>
</tr>
<tr>
<td>少林サッカー</td><td>座頭市</td>
<tr>
<td>アメリ</td><td>パイレーツ・オブ・カリビアン</td>
</tr>
<tr>
<td>ビューティフル・マインド</td>
<td><font color='red'>ロボコン</font></td>
</tr>
</table>
</body>
</html>
```

図 2.3: テーブル例の HTML

図 2.3 のタグには、テーブルの枠線の太さを指定する 'border' や背景色を指定する 'bgcolor' などの属性が記述されているが、これらの属性は無視する。コンテンツを修飾する <font> タグなども無視する。

さらに、複数行にわたるコンテンツは抽出しない。ウェブにおいてテーブルは様々な場面で使用されており、例えば文章を整形して表示するために <table> タグが用いられることがある。複数行にわたるコンテンツは固有表現である可能性が低いので無視する。

また、あらかじめストップワードを設定しておけば、ストップワードにマッチするコンテンツを取り除くことができる。例えば、料理名の抽出を行いたいときに、料理名が多く現れるテーブルに、その料理の値段も現れることがある。このとき、『'数字'+'円'』をストップワードに設定しておけば、これらの値段は抽出対象から排除できる。

図 2.3 の HTML ファイルから抽出されるテーブルを図 2.4 に示す。<table>、</table> タグに囲まれた、<table> タグ、<tr> タグ、<td> タグおよびコンテンツのみが抽出される。『踊る大捜査線 THE MOVIE 2 レインボーブリッジを封鎖せよ!』というコンテンツは 2 行にわたっているため抽出されない。

2.2.2 コンテンツの抽出と番号の付与

次に、テーブルからコンテンツのみを抽出する。このとき、シードワードと同列の固有表現を抽出するための前準備として、コンテンツにテーブル番号、行番号、および列番号を付与する。番号は以下のルールにしたがって付与される。

- <table> タグが現れればテーブル番号を 1 増加させ、行番号と列番号は 0 に戻す。

<table>
<tr>
<td>
2002年に観た映画
<td>
2003年に観た映画
<tr>
<td>
モンスターズインク
<td>
<tr>
<td>
少林サッカー
<td>
座頭市
<tr>
<td>
アメリ
<td>
パイレーツ・オブ・カリビアン
<tr>
<td>
ビューティフル・マインド
<td>
ロボコン

図 2.4: 抽出されるテーブル

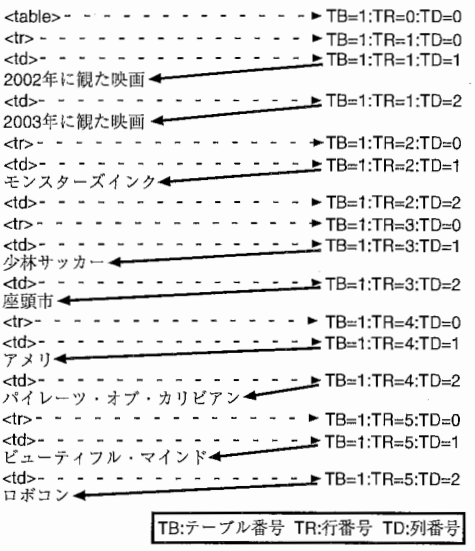


図 2.5: 番号の付与

- <tr> タグが現れれば行番号を 1 増加させ、列番号は 0 に戻す。
- <td> タグが現れれば列番号を 1 増加させる。
- コンテンツが現れればその時点でのテーブル番号、行番号、列番号をコンテンツに付与する。

図 2.4 のテーブルのコンテンツに対する番号付けの様子を図 2.5 に示す。図において、TB はテーブル番号、TR は行番号、TD は列番号である。

2.2.3 シードワードと同列の固有表現の抽出

最後に、固有表現抽出部ではシードワードと同じテーブルで同じ列に現れるコンテンツを固有表現として抽出する。コンテンツにはテーブル番号、行番号、列番号が付与されているので、シードワードと同じテーブルで同じ列に現れるコンテンツを抽出するには、シードワードとテーブル番号および列番号が等しいコンテンツを抽出すれば良い。ただし、テーブルの 1 行目は見出しである可能性があるので、行番号が '1' であるコンテンツは抽出しない。

例えば、『アメリ』をシードワードとしたときに図 2 のようなテーブルを持つウェブページが検索されれば、図 5 より『アメリ』のテーブル番号、行番号、列番号はそれぞれ '1'、'4'、'1' なので、このテーブルからはテーブル番号と列番号が '1' であり、なおかつ行番号が '1' ではない『モンスターズインク』、『少林サッカー』および『ビューティフル・マインド』が抽出される。

2.3 確信度付与部

最後に、固有表現抽出システムでは抽出されたシードワードと同列の固有表現に確信度を付与する。この確信度は、以下の考え方に基づいて決定される。

- スコアの高いテーブルから抽出された固有表現の確信度は高い。
- 確信度の高い固有表現が現れるテーブルのスコアは高い。

このような考え方に基づき、固有表現 e_i の確信度 C_{e_i} とテーブル t_m のスコア S_{t_m} を以下のように定義する。

$$C_{e_i} = \sum_j S_{t_j} \quad \{t_j | t_j \in \mathrm{T}_{e_i}\} \tag{2.1}$$

$$S_{t_m} = \frac{\sum_n C_{e_n}}{N_{t_m}} \quad \{e_n | e_n \in \mathrm{E}_{t_m}\} \tag{2.2}$$

ここで、T_{e_i} は固有表現 e_i が現れるテーブルの集合、E_{t_m} はテーブル t_m から抽出される固有表現の集合、N_{t_m} はテーブル t_m から抽出される固有表現の総数である。

(2.2) 式の右辺をそのテーブルから抽出される固有表現の総数 N_{t_m} で割っているのは、確信度の高い固有表現を高い割合で抽出しているテーブルのスコアが高くなるようにするためである。

(2.1)、(2.2) 式を見ると、固有表現の確信度を表すのにテーブルのスコアが用いられ、テーブルのスコアを表すのに固有表現の確信度が用いられている。この式に基づいて、固有表現の確信度は以下のようにして決定される。

1. それぞれの固有表現が現れるテーブル数をその固有表現の確信度の初期値とする。
2. (2.2) 式に基づきテーブルのスコアを求める。
3. (2.1) 式に基づき固有表現の確信度を求める。
4. 2. と 3. のステップを固有表現の確信度の値が収束するまで繰り返す。

ただし、固有表現の確信度の値を収束させるために、固有表現の確信度は必ず以下の式を満たすように正規化する。

$$\sum_i C_{e_i} = 1 \tag{2.3}$$

2.4 固有表現抽出の繰り返し

固有表現抽出システムは、シードワードとなる数語の固有表現を入力すると、シードワードと同ジャンルの固有表現を確信度で順序付けて出力する。この一連の固有表現の抽

出を 1 サイクルと呼ぶこととする。1 サイクル毎に確信度付き固有表現のリストが出力されるので、固有表現の抽出を繰り返す際には、それぞれのサイクルで出力される確信度に基づき、全体を通じての確信度を決定する必要がある。

本システムでは、より多くのサイクルで抽出された固有表現の確信度は高いという考えに基づき、固有表現 e_i の全サイクルを通じての確信度 G_{e_i} を以下のように定義する。

$$G_{e_i} = F_{e_i} \times \sum_r C_{e_i,r} \tag{2.4}$$

ここで、F_{e_i} は全サイクル中 e_i が抽出されたサイクル数、$C_{e_i,r}$ は r サイクル目における e_i の確信度である。

第3章

実験

3.1 固有表現抽出実験

作成した固有表現抽出システムで、「映画のタイトル」、「空港名」、「ホテル名」、「京都の観光地」および「料理名」という5つのジャンルの固有表現の抽出実験を行った。それぞれ、固有表現の抽出を50サイクル繰り返し、50サイクル終了時点での確信度上位100位の表現を以下の3段階で評価した。

正解　シードワードと同ジャンルの固有表現である。
半正解　シードワードと同ジャンルの固有表現を部分的に含む。
不正解　シードワードと同ジャンルの固有表現ではない。

それぞれのジャンルでシステムに入力として与えたシードワードを表3.1に、抽出を50サイクル繰り返した時点での確信度上位100位の評価結果を表3.2に示す。

3.2 考察

表3.2より、50サイクル終了時点での確信度上位100位の表現の内、正しく抽出された固有表現が90個以上存在した。

一方、正しく抽出されなかったものとしては、例えば映画のタイトルの抽出で「エスプレッソマシーン」や「コーヒー豆の種類」といった映画とは関係のない表現が抽出された。

今回のシステムでは、単純にシードワードと同じテーブルで同じ列に現れる表現を固有表現として抽出しているので、シードワードと同列に正しくない表現が存在すればその表現は抽出されてしまう。しかし、より多くのテーブルやサイクルから抽出された固有表現の確信度が高くなるように確信度を定義したので、あまり多くは抽出されないであろう正しくない表現の確信度は低くなるはずである。

表 3.1: システムに与えたシードワード

ジャンル	シードワード
映画のタイトル	モンスターズインク、アメリ、少林サッカー
空港名	関西国際空港、新東京国際空港、那覇空港、新千歳空港
ホテル名	京都ロイヤルホテル、ウェスティン都ホテル京都、からすま京都ホテル
京都の観光地	平安神宮、京都御所、貴船神社、常寂光寺
料理名	プレーンオムレツ、ハンバーグ、ビーフカレー

表 3.2: 50 サイクル終了時点での確信度上位 100 位の評価結果

ジャンル	正解数	半正解数	不正解数	計
映画のタイトル	89	0	11	100
空港名	97	0	3	100
ホテル名	94	0	6	100
京都の観光地	100	0	0	100
料理名	91	2	7	100
計	471	2	27	500

にもかかわらず、今回このような表現が確信度上位に抽出されたのは、これらの表現が同一のフレームを持つ複数のウェブページから抽出されたためである。図 3.1 に、これらの正しくない表現が実際に抽出されたウェブページを示す。これらのページのフレーム部分では、映画のタイトルと同じテーブルの同じ列に「エスプレッソマシーン」や「コーヒー豆の種類」といった言葉が含まれており、なおかつこのフレームが複数のウェブページで共通して使用されている。よって、このフレームを含むウェブページを複数ダウンロードした場合、システムでは、異なるページから抽出されたテーブルは違うテーブルであると判断するため、これらの正しくない表現が多くのテーブルから抽出されたとみなされ、確信度が大きくなってしまう。

この問題を防ぐ方法として、同じドメインで同じホームディレクトリ名を持つウェブページからは 1 度しかダウンロードしないようにするという方法が考えられる。フレームが共通するページは同じドメインで同じホームディレクトリ名を持つウェブページに現れる場合が多いからである。

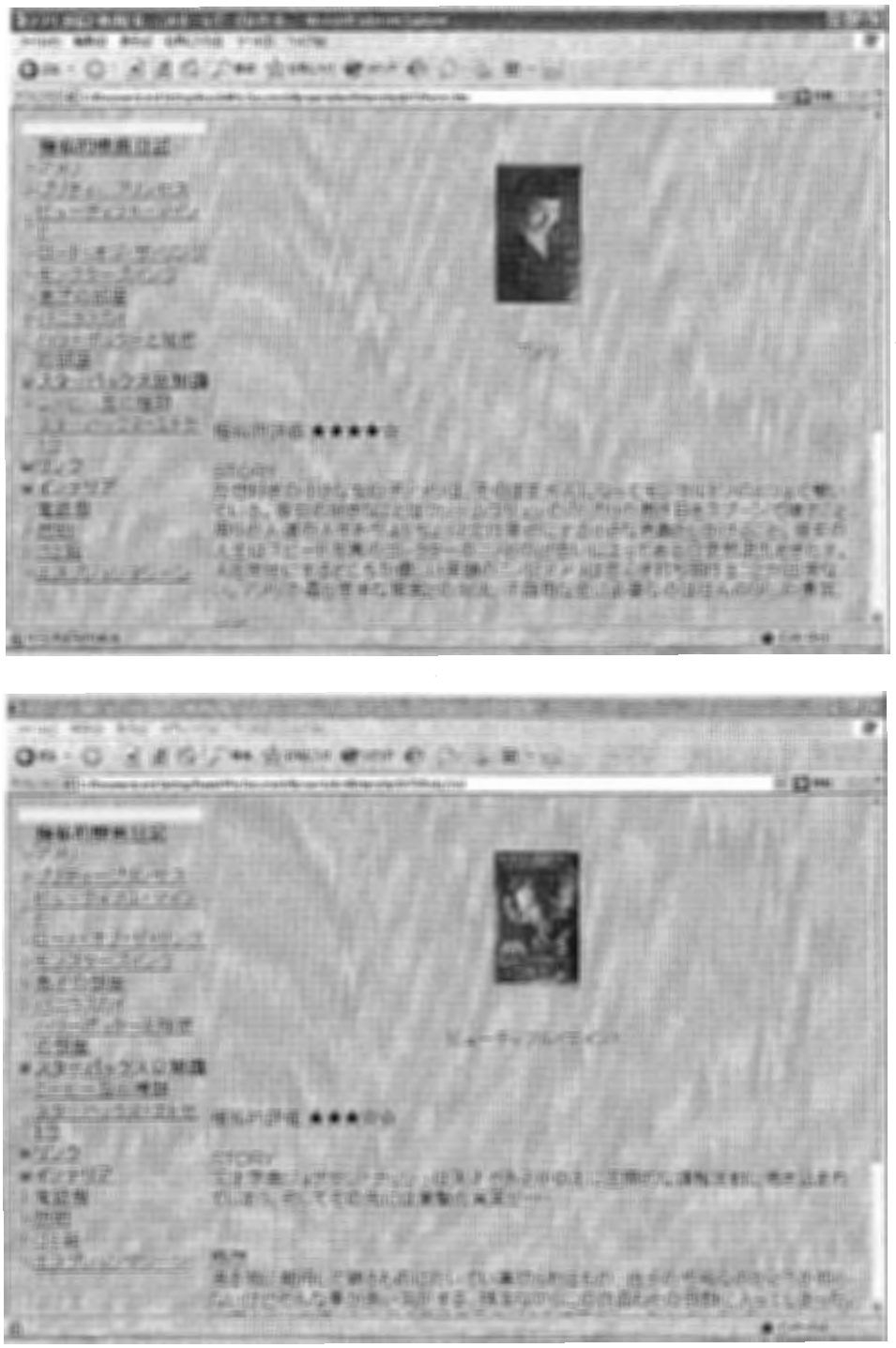


図 3.1: 同じフレームから誤った表現が抽出されてしまう例

第4章

さらなる改善に向けて

4.1 ユーザへの問い合わせ

今回の実験の範囲では、50サイクル終了時点での確信度100位以上では正しい固有表現を9割以上の精度で抽出することができたが、さらに抽出を繰り返すと徐々に誤った表現が多く抽出されるようになることがある。

例えば、空港名の抽出を100サイクル繰り返した時点での確信度上位100位には航空会社名が39個含まれる結果となった。今回の実験では、確信度上位の表現の内、それまでシードワードとして用いられていない表現3個をシステムの新たな入力として固有表現の抽出を繰り返した。空港名を抽出する際にシステムに与えられたシードワードを見ると、50サイクル目まではほぼ空港名がシードワードとして与えられているが、50サイクル目からは地名がシードワードとなり、62サイクル目からはほとんどが航空会社名となっている(表4.1)。これは、日本の主要な空港名が50サイクル目までのシードワードにほぼ出尽くしてしまったことが原因であると思われる。

この対策として、抽出を決められた回数繰り返したら、ユーザにそれまでに与えられたシードワードは適切か問い合わせる、もしくは確信度上位に正しい固有表現が抽出できているかを問い合わせる、という方法が考えられる。ユーザからの解答に基づいて、確信度を計算し直したり、あるいは十分な数の固有表現が抽出できていればそこでシステムを終了させることができる。

4.2 内部構造に基づく確信度

固有表現には、名前を特徴づける要素を持つものがある。例えば、料理名を特徴づける要素としては料理の素材名や調理法などがあり、『'素材名'の'調理法'』という構造を持った料理名が数多く存在する。料理名の抽出を行う際には、このような構造を持つ表現の確

表 4.1: 空港名の抽出時に与えられたシードワード

サイクル	シードワード
1	関西国際空港、新東京国際空港、那覇空港、新千歳空港
2	鹿児島空港、仙台空港、名古屋空港
3	小松空港、長崎空港、松山空港
4	高知空港、宮崎空港、岡山空港
5	大分空港、熊本空港、高松空港
6	米子空港、福岡空港、新潟空港
7	広島空港、佐賀空港、出雲空港
8	富山空港、成田空港、福島空港
9	徳島空港、松本空港、山口宇部空港
10	鳥取空港、南紀白浜空港、羽田空港
11	大館能代空港、広島西空港、函館空港
12	花巻空港、庄内空港、秋田空港
13	青森空港、山形空港、三沢空港
14	対馬空港、旭川空港、釧路空港
15	隠岐空港、帯広空港、稚内空港
16	女満別空港、但馬空港、北九州空港
17	八丈島空港、石見空港、大島空港
18	三宅島空港、宮古空港、石垣空港
19	奥尻空港、利尻空港、札文空港
20	オホーツク紋別空港、中標津空港、静岡空港
21	空港の概要、中部国際空港、徳之島空港
22	丘珠空港、屋久島空港、沖永良部空港
23	壱岐空港、喜界空港、広島西飛行場
24	波照間空港、久米島空港、北大東空港
25	多良間空港、種子島空港、粟国空港
26	慶良間空港、下地島空港、与那国空港
27	南大東空港、与論空港、伊江島空港
28	奄美空港、紋別空港、新島空港
29	福江空港、小値賀空港、上五島空港
30	五島福江空港、佐渡空港、福井空港
31	空港名、大阪国際空港、神戸空港
32	能登空港、奄美大島空港、八尾空港
33	びわこ空港、空港ビル、バス
34	神津島空港、有明佐賀空港、伊丹空港
35	枕崎空港、乗入航空会社は変更になる場合があります。、東京国際空港
36	根室中標津空港、鉄道、関西空港
37	大阪空港、天草空港、喜界島空港
38	新東京国際空港（成田）、とかち帯広空港、新北九州空港
39	岡南飛行場、東京国際空港（羽田）、コウノトリ但馬空港
40	米軍横田基地、北海道、東京ヘリポート
41	千歳飛行場、弟子屈飛行場、関東
42	東北、沖縄、空港一般
43	四国、中国、九州
44	購入先、タイプ、岡山
45	中部、近畿、百里飛行場
46	-、徳島飛行場、大阪国際空港（伊丹）
47	栗国空港、三沢飛行場、無し
48	空港の案内、米軍嘉手納基地、米軍三沢基地
49	出雲空港（AIR-WEB）、米子空港（AIR-WEB）、空港の気象
50	飛騨エアパーク、船、高知
51	山形、青森、長崎
52	福岡、鳥取、富山
53	福島、新潟、広島
54	仙台、奄美、三宅島
55	八丈島、米子、大島
56	大分、宮崎、鹿児島
57	佐賀、熊本、名古屋
58	徳島、秋田、北九州
59	計、大館能代、小松
60	那覇、松山、空港
61	ソウル、羽田、広島西
62	日本航空、コリアンエアー、アシアナ航空
63	ノースウエスト航空、エアージャパン、日本エアシステム
64	ユナイテッド航空、出雲、屋久島
65	中国国際航空、中国南方航空、ニュージーランド航空
66	フィリピン航空、シンガポール航空、タイ国際航空
67	マレーシア航空、大韓航空、ベトナム航空
68	ガルーダ・インドネシア航空、チャイナエアライン、アリタリア航空
69	トルコ航空、デルタ航空、中国東方航空
70	コンチネンタル航空、スカンジナビア航空、キャセイパシフィック航空
71	ヴァリグ・ブラジル航空、エア・インディア、KLM オランダ航空
72	アメリカン航空、オーストリア航空、ブリティッシュ・エアウェイズ
73	フィンランド航空、エバー航空、エジプト航空
74	イラン航空、日本アジア航空、カナディアン航空
75	エア・カナダ、パキスタン航空、スイス航空
76	ヴァージンアトランティック航空、全日空、エアーニッポン
77	中国西北航空、中国北方航空、ルフトハンザ航空
78	サベナ・ベルギー航空、エールフランス航空、ルフトハンザドイツ航空
79	KLMオランダ航空、ルフトハンザ・ドイツ航空、エアランカ航空
80	コンチネンタル・ミクロネシア航空、ニューギニア航空、カンタスオーストラリア航空
81	航空会社、全日本空輸、中国西南航空
82	中華航空、香港ドラゴン航空、カンタス・オーストラリア航空
83	エールフランス、モンゴル航空、アエロフロート・ロシア航空
84	ハワイアン航空、ウズベキスタン航空、ロイヤルネパール航空
85	ガルーダインドネシア航空、メキシカーナ航空、アロハ航空
86	スリランカ航空、ノースウェスト航空、エア・パシフィック
87	パキスタン国際航空、JAL ウェイズ、エールフランス国営航空
88	アンセット・オーストラリア航空、カンタス航空、南アフリカ航空
89	オリンピック航空、日本トランスオーシャン航空、英国航空
90	コンチネンタルミクロネシア航空、サウスウエスト航空、アメリカウエスト航空
91	エアインディア、ビーマン・バングラデシュ航空、ヴァリグブラジル航空
92	アラスカ航空、サハリン航空、アエロフロート・ロシア国際航空
93	モーリシャス航空、エミレーツ航空、↑ BACK
94	サウジアラビア航空、ウラジオストク航空、チェコ航空
95	搭乗便名、MIAT モンゴル航空、ドラゴン航空
96	キャセイ・パシフィック航空、サベナベルギー航空、イベリア航空
97	出発時間、フロンティア航空、出 発 日
98	ビーマン・バングラディッシュ航空、アフリカウエスト航空、イベリア・スペイン航空
99	隠岐、但馬、トランスワールド航空
100	南紀白浜、東京、大阪

信度を高くすると抽出精度が向上する可能性がある。

第5章

まとめ

本研究では、音声認識のための語彙を獲得する方法として、ウェブ上のテーブルから固有表現を抽出する方法を提案した。本研究で作成した固有表現抽出システムでは、シードワードとして抽出したいジャンルの固有表現を数語入力すると、シードワードと同ジャンルの固有表現を確信度付きで出力する。確信度上位の固有表現をシステムの新たな入力とすることにより、繰り返し固有表現を抽出することが可能であることを確認した。