概要

ブログからは，趣味性が高く詳しい情報が得られ，特にどこにどんな物があるかという情報は旅行を盛り上げる材料になりうる．観光支援として存在性情報（どこに何があるか）の抽出が行われている．先行研究 [1] では，ブログ記事からパターン対を用いた場所と存在物の情報抽出が行われた．

ここで，存在物や存在場所の抽出は固有表現抽出 [2] の一種と考えられる．存在情報の抽出と固有表現抽出の差は，一般名詞による存在物や場所の表現を抽出しなければならないこと，および，存在物と存在場所の対応を検出しなければならないことである．そこで本研究では，柔軟性を持った手法として SVM[3] を用いて，文章から存在物と場所の抽出，および，それらの対応を検出することを提案する．

具体的には，まず，タグ付きコーパスを作成する．固有表現抽出のタスクにはタグ付きのコーパスが必要になる．ブログ記事から「ドクターイエロー」に関係するブログ記事を抽出し，構文解析を行う．解析結果に，存在物および場所の表現に IOB2 タグ [4] を人手で付ける．また，存在物に ID を付与し, 存在する場所に存在物 ID を「存在物リンク」として付与する．抽出は工藤らの手法を利用する．

次に，存在物と場所の対応を検出する．ベースライン手法は存在物と場所の単語間の距離が一番近いものを採択する．提案手法は存在物一つに対して記事内全ての場所とそれぞれペアにし、各ペアの存在物と場所が対応しているか SVM に判定させる手法である．素性は 13 種類あり，組み合わせによって 16 種類の実験を行う．

２つの手法の実験結果について，抽出結果と正解データの F 値で手法の評価を行う．評価方法には，SVM のスコアが正値かつ最大値のペアを推定結果とする方法 M_{sgx}，および，正値のペアをすべて推定結果とする方法 M_{plx} を設ける $(x = 1, 2, ..., 16)$．リンク単位での対応検出の結果，ベースライン手法では，F 値が 0.30 となった．提案手法では，M_{sg11} の F 値は 0.24，M_{pl11} は 0.52 となった.「ドクターイエロー」に関する対応検出の結果，ベースライン手法では，F 値が 0.67 となった．提案手法では，M_{sg11} の F 値は 0.56，M_{pl11} は 0.60 となった．ベースライン手法の F 値を越えることはでなかった．

コーパス依存性を確認するため，コーパスをドクターイエローコーパスからお土産コー

パスに変更して実験を行った.「赤福」に関する対応検出の結果，ベースライン手法では，F 値が 0.61 となった．提案手法では，M_{sg12} の F 値は 0.61，M_{pl12} は 0.65 となり，F 値の向上が確認できた.

さらに，一般的な方法と比較を行う．普段，存在性情報を得る時は本やインターネットを利用する．そこで，Google 検索の結果と提案手法の比較を行った．ドクターイエローコーパスの場合，Google 検索で得ることができた存在する場所は駅名がほとんどであった．しかし，提案手法では駅名の他にも，富士川や中里などの存在する場所も得ることができた.

お土産コーパスの実験で F 値の向上を確認できたこと，Google 検索との比較で Google 検索で得られない場所を得られたことから，提案手法に対する一定の評価を得ることができたと考える．今後の課題は，場所から存在物の対応検出を行うこと，および，時間の存在する時間（いつ見ることができるか）の情報抽出を行うことである.

目 次

図 目 次

表 目 次

第1章　はじめに

ブログからは，趣味性が高く詳しい情報が得られ，特にどこにどんな物があるかという情報は旅行を盛り上げる材料になりうる．観光支援として存在性情報（どこに何があるか）の抽出が行われている．先行研究 [1] では，ブログ記事からパターン対を用いた場所と存在物の情報抽出が行われた．

ここで，存在物や存在場所の抽出は固有表現抽出 [2] の一種と考えられる．存在情報の抽出と固有表現抽出の差は，一般名詞による存在物や場所の表現を抽出しなければならないこと，および，存在物と存在場所の対応を検出しなければならないことである．

そこで本研究では，SVM[3] を用いて，文章から存在物と場所の抽出，および，それらの対応を検出することを目的とする．

第 2 章では，関連研究について述べる．第 3 章では，コーパスの作成について述べる．第 4 章では，存在物の抽出と場所の抽出について述べる．第 5 章では，存在物と場所の対応検出について述べる．第 6 章では，コーパスを変更して対応検出を行う追加実験について述べる．第 7 章では，オープンテストついて述べる．第 8 章では，考察を行う．第 9 章では，まとめを行う．

第2章　関連研究

本章では，本研究で用いる自然言語処理（NLP）の基本技術を解説した後，存在性情報抽出のタスクについて説明する．

2.1　基本技術

2.1.1　シソーラス

シソーラスとは言語を同意義語や意味上の類似関係，包含関係などによって分類した辞書である．この辞書には，分類語彙表，日本語語彙大系 [5]，日本語大シソーラス等がある．本研究で用いる日本語語彙大系は，日本語の語彙 30 万語を 3,000 種類の意味属性で分類したシソーラスである．意味体系，単語体系，および，構文体系の 3 部から構成されている．意味属性体系には一般名詞意味属性体（2,700 属性），固有表現名詞意味属性体系（130 属性），および，文型パターン対に対する用言意味属性体系（100 属性）のうちの上位 36 属性の各大系と，各意味属性別の単語表が収録されている [6]．一般名詞意味属性および用言意味属性のコードを取得し，NI コード，および，NY コードとして本研究で利用する．

2.1.2　形態素・構文解析

形態素解析は文章を意味のある単語に区切り，辞書を利用して品詞や内容を判別することである．ソフトとしては ChaSen や MeCab などがある．構文解析は，文節間の係り受け構造を発見することである．ソフトとしては KNP や CaboCha[7] などがある．本実験では CaboCha を用いる．CaboCha は SVM に基づく日本語係り受け解析器である．入力文から単語境界，品詞，固有表現タグ，および，係り関係の付与された情報を得ることができる．固有表現については 2.3 で述べる．例文「理由はもちろんドクターイエローが走るから。」を解析した結果を図 2.1 に示す．

```
<sentence>
<chunk id="0" link="3" rel="D" score="4.43075" head="0" func="1">
<tok id="0" read="リユウ" base="理由" pos="名詞-一般" ctype=""
cform="" ne="O">理由 </tok>
<tok id="1" read="ハ" base="は" pos="助詞-係助詞" ctype="" cform=""
ne="O">は </tok>
</chunk>
<chunk id="1" link="3" rel="D" score="3.83794" head="2" func="2">
<tok id="2" read="モチロン" base="もちろん" pos="副詞-一般" ctype=""
cform="" ne="O">もちろん </tok>
</chunk>
<chunk id="2" link="3" rel="D" score="0" head="4" func="5">
<tok id="3" read="ドクター" base="ドクター" pos="名詞-一般" ctype=""
cform="" ne="O">ドクター </tok>
<tok id="4" read="イエロー" base="イエロー" pos="名詞-一般" ctype=""
cform="" ne="O">イエロー </tok>
<tok id="5" read="ガ" base="が" pos="助詞-格助詞-一般" ctype=""
cform="" ne="O">が </tok>
</chunk>
<chunk id="3" link="-1" rel="O" score="0" head="6" func="7">
<tok id="6" read="ハシル" base="走る" pos="動詞-自立"
ctype="五段・ラ行" cform="基本形" ne="O">走る </tok>
<tok id="7" read="カラ" base="から" pos="助詞-接続助詞" ctype=""
cform="" ne="O">から </tok>
<tok id="8" read="。" base="。" pos="記号-句点" ctype="" cform=""
ne="O">。</tok>
</chunk>
</sentence>
```

図 2.1: 係り受け解析結果の例

2.1.3 固有表現抽出

固有表現抽出とは，情報検索，情報抽出の基礎として，テキスト中から人名，地名，組織名などを自動的に抽出を行う処理である．SVM[3] や CRF を用いた機械学習に基づく手法で高い精度が報告されている．とくに SVM では文頭または文末から決定的に固有表現タグを決定していく，系列ラベリングを用いた方法で高い精度に達成している．

固有表現抽出の先行研究については 2.5.2 で述べる．

2.1.4 機械学習とチャンキング問題

機械学習には教師あり学習，教師なし学習，半教師あり学習などがある．教師あり機械学習は事前に与えられたデータを学習し，未知のデータを与えたときに学習データを元に分類する方法である．SVM は教師あり機械学習を用いる識別手法のひとつである．

チャンキング（任意句の同定）問題に学習手法として用いている [4]．この問題を解くにはタグ付きコーパスが必要となる．本研究では IOB2 タグをコーパスに用いる．IOB2 タグはチャンク（任意句）の状態を表すタグである．I はチャンクの内部，O はチャンクの外部，B はチャンクの開始地点を表す．日本語の文法は SVO で構成されているので，文末から文頭の順に推定を行う．よって推定を行う際は，IOB2 タグを IOE2 タグに変換する．IOE2 タグの I はチャンクの内部，O はチャンクの外部，E はチャンクの終了地点を表す．

2.2 存在性情報の抽出タスク

2.2.1 パターン対を用いた存在物と場所の抽出

北尾らはパターン辞書を用いて，2 文から存在性情報を得た [1]．例えば「名古屋駅に行きました。ドクターイエローを見ました。」という入力文がある．パターン辞書には FP(first pattern) に「$N1$ に行く」，SP(second pattern) に「$N1$ を見る」というパターンがあったとする．1 文目は SP に，2 文目は SP と適合する．よって場所「名古屋駅」，存在物「ドクターイエロー」を抽出し，機械的に「名古屋駅にドクターイエローが存在する」ことを検出できるようにした．

北尾らの研究では１文に対し必ず１ヶ所動詞がないとパターンを用いることができない．例えば「名古屋駅に行きました。ドクターイエローです。」という入力文がある．先

行研究のパターン適合の方法では，1 文目は SP に適合するが，2 文目は SP と適合しない．よって,「名古屋駅にドクターイエローが存在する」ことを検出でない．

2.2.2 日本語固有表現認識

笹野らは SVM を用いて IREX[8] で定義された固有表現の抽出を行った [2]．抽出する表現のタイプとしては人名，地名，組織名などの固有名詞的表現のほかにも時間表現や数値表現を対象とした．笹野らが抽出した固有表現を表 2.1 に示す．

表 2.1: 笹野らが抽出した固有表現の種類と例

	固有表現の種類		例
固有名詞的表現	組織名	ORGANIZATION	NHK 交響楽団，ICAO
	人名	PERSON	福田康夫，川崎憲次郎
	地名	LOCATION	アメリカ，新義州
	人口物名	ARTIFACT	ノーベル賞，ひかり 123 号
時間表現	日付	DATE	6 月 17 日，今年
	時刻	TIME	午後五時，正午
数値表現	金額	MONEY	500 円，五・七新ペソ
	割合	PERCENT	90%，三分の一

笹野らの研究では一般名詞は抽出の対象としていない．また，抽出した物の存在性情報は解析していない．

2.3 本研究の位置づけ

本研究では，観光支援として存在性情報（どこに何があるか）の抽出が目的である．まず，チャンキング問題として存在物の抽出と場所の抽出を行う．抽出は一般名詞の抽出も可能にすることを目指す．次に，柔軟性を持たせる方法として SVM を用いて存在物と場所の対応検出を行う．コーパスはドクターイエローコーパス，および，お土産コーパスを作成する．

第3章　コーパスの作成

本研究での，コーパスの作成方法を説明する．コーパスに存在物，場所，および，存在する場所についての情報を明示することを目的とする．また，機械学習の素性の基となる情報と対応付けも行う．

3.1　手順

手順１： ブログから「ドクターイエロー」に関する記事を抽出する．

抽出する記事の条件は，記事内に存在物と場所がそれぞれ 1 表現以上あるもの，および，1 つ以上存在物が存在する場所があることとする．

手順２： 記事内の文を CaboCha で構文解析し，単語境界，品詞，固有表現タグ，係り先の情報を得る．

手順３： 存在物および場所の表現に IOB2 タグを人手で付ける．

本研究では存在物と場所の定義を以下のようにする．

存在物： 車両（ドクターイエロー，新幹線など），食品（駅弁など），展示品（銅像，仏像など），おみやげのような存在する具体物．

場所： 固有表現タグ LOCATION タグがあるもの（地名），さらに，自然のもの（山，川など），建築物（駅，道路，橋など）よのうな移動することがない存在物．

場所にタグがつくものは存在物のタグを付与していない．

手順４： 存在物に ID を付与し，存在する場所に存在物 ID を「存在物リンク」として付与する．1 つの場所に複数の存在物がある場合，複数の存在物 ID を付与する．存在物 ID は存在物タグの B，存在物リンクは場所タグの B が付与された単語に付与する．存在物 ID は記事単位でユニークとする．

ただし，存在する場所にも制限を設ける．今回の目的は観光支援であるため，存在物を買うことができる，鑑賞することができる，食べることができるなど，観光に有益な場所のみを存在する場所とする．

例えば,「鳥取に帰って来ました。名古屋駅で赤福を買ったので，今から食べます。」という記事がある．赤福は鳥取に存在しているが，鳥取で赤福が買えるかどうかはこの記事からわからない．この場合，赤福の存在物リンクが付与されるのは名古屋駅のみとなる．

3.2 コーパスの例

人手でタグ付けを行った部分の例を表 3.1 に示す．例文は「名古屋駅で N700 系とドクターイエローを撮影しました。」である．存在物は「Ｎ７００系」と「ドクターイエロー」である．場所は「名古屋駅」である.「Ｎ７００系」は「名古屋駅」に存在する．また,「ドクターイエロー」も「名古屋駅」に存在する．

表 3.1: タグ付けの例

単語	存在物タグ	存在物 ID	場所タグ	存在物リンク
名古屋	O		B	1;2
駅	O		I	
で	O		O	
N	B	1	O	
7	I		O	
0	I		O	
0	I		O	
系	I		O	
と	O		O	
ドクター	B	2	O	
イエロー	I		O	
を	O		O	
撮影	O		O	
し	O		O	
まし	O		O	
た	O		O	
。	O		O	

3.3 結果

2013 年 2 月〜4 月のブログからドクターイエローに関係する記事は 84 記事抽出された．文数は 1,507，単語数は 24,499 となり，存在物は 566 箇所，場所は 458 箇所あった．存在物についてのタグは，B が 566，I が 983 で，場所についてのタグは B が 458，I が 421 になった．存在物リンクの付与された場所は 345ヶ所であった．存在物と場所のリンク数は 2,240 であった．対応する場所の無い存在物は 41 件であった．

84 記事のうち 20 記事は記事内に場所が 1 箇所しかなかった．リンク数にすると 63 であった．そのうち存在物がドクターイエローのペアは 24 あり，20 ペアは存在する場所のペアであった．

第4章　存在物の抽出と場所の抽出

存在物の抽出と場所の抽出をそれぞれ行う．チャンキング問題としてSVMを用いて抽出を行う．

4.1　手法

4.1.1　ベースライン手法

固有表現タグで抽出することをベースライン手法 (B_1) にする．場所はLOCATIONタグが付く単語，存在物はARTIFACTタグが付く単語とする．

4.1.2　提案手法

SVMを用いて文末から文頭の順に各単語のIOB2タグを推定する．素性は，次の単語，品詞，固有表現タグ，係り先の情報，および，次の単語の推定IOB2タグとする．係り先の情報は現在の単語とその先の単語を組み合わせた単語列とする．

4.2 実験の様子

4.2.1 手順

手順 1： コーパスを分割する

コーパスを 8 分割し，そのうち 1 つをテストデータ，他を学習データとする．

手順 2： IOB2 タグを IOE2 タグに変換する．存在物タグ部分の変換例を表 4.1 に示す．

表 4.1: タグ変換の例

単語	変換前	変換後
名古屋	O	O
駅	O	O
で	O	O
N	B	I
7	I	I
0	I	I
0	I	I
系	I	E
と	O	O
ドクター	B	I
イエロー	I	E
を	O	O
撮影	O	O
し	O	O
まし	O	O
た	O	O
。	O	O

手順 3： 学習データの素性を作成

学習データに 4.1.2 節で述べた素性を作成する．入力文「名古屋駅でＮ７００系とドクターイエローを撮影しました。」の存在物の抽出実験時の例を図 4.1 に示す．

```
0 O NXT:CLS:O NXT:駅 pos:名詞-固有名詞-地域-一般 ne:B-LOCATION DP:nil
1 O NXT:CLS:O NXT:で pos:名詞-接尾-地域 ne:I-LOCATION DP:nil
2 O NXT:CLS:I NXT:Ｎ pos:助詞-格助詞-一般 ne:O DP:で撮影
3 O NXT:CLS:I NXT:７ pos:記号-アルファベット ne:O DP:nil
4 O NXT:CLS:I NXT:０ pos:名詞-数 ne:O DP:nil
5 O NXT:CLS:I NXT:０ pos:名詞-数 ne:O DP:nil
6 O NXT:CLS:E NXT:系 pos:名詞-数 ne:O DP:nil
7 O NXT:CLS:O NXT:と pos:名詞-接尾-一般 ne:O DP:nil
8 O NXT:CLS:I NXT:ドクター pos:助詞-並立助詞 ne:O DP:とドクター
9 O NXT:CLS:E NXT:イエロー pos:名詞-一般 ne:O DP:nil
10 O NXT:CLS:O NXT:を pos:名詞-一般 ne:O DP:nil
11 O NXT:CLS:O NXT:撮影 pos:助詞-格助詞-一般 ne:O DP:を撮影
12 O NXT:CLS:O NXT:し pos:名詞-サ変接続 ne:O DP:nil
13 O NXT:CLS:O NXT:まし pos:動詞-自立 ne:O DP:nil
14 O NXT:CLS:O NXT:た pos:助動詞 ne:O DP:nil
15 O NXT:CLS:O NXT:。 pos:助動詞 ne:O DP:nil
16 O NXT:CLS:X NXT:EOS pos:記号-句点 ne:O DP:nil
```

図 4.1: 学習データの例

手順 4： SVM に学習データを学習させる．

手順 5： テストデータの素性を作成

テストデータに 4.1.2 節で述べた素性を元に作成する．ただし次のクラスが X，I，O，および，E だけのものをそれぞれ作成する．

入力文「名古屋駅でＮ７００系とドクターイエローを撮影しました。」の存在物の抽出実験時の例を図 4.2 に示す．

```
0 O NXT:CLS:X NXT:駅 pos:名詞-固有名詞-地域-一般 ne:B-LOCATION DP:nil
1 O NXT:CLS:X NXT:で pos:名詞-接尾-地域 ne:I-LOCATION DP:nil
2 O NXT:CLS:X NXT:Ｎ pos:助詞-格助詞-一般 ne:O DP:で撮影
3 O NXT:CLS:X NXT:７ pos:記号-アルファベット ne:O DP:nil
4 O NXT:CLS:X NXT:０ pos:名詞-数 ne:O DP:nil
5 O NXT:CLS:X NXT:０ pos:名詞-数 ne:O DP:nil
6 O NXT:CLS:X NXT:系 pos:名詞-数 ne:O DP:nil
7 O NXT:CLS:X NXT:と pos:名詞-接尾-一般 ne:O DP:nil
8 O NXT:CLS:X NXT:ドクター pos:助詞-並立助詞 ne:O DP:とドクター
9 O NXT:CLS:X NXT:イエロー pos:名詞-一般 ne:O DP:nil
10 O NXT:CLS:X NXT:を pos:名詞-一般 ne:O DP:nil
11 O NXT:CLS:X NXT:撮影 pos:助詞-格助詞-一般 ne:O DP:を撮影
12 O NXT:CLS:X NXT:し pos:名詞-サ変接続 ne:O DP:nil
13 O NXT:CLS:X NXT:まし pos:動詞-自立 ne:O DP:nil
14 O NXT:CLS:X NXT:た pos:助動詞 ne:O DP:nil
15 O NXT:CLS:X NXT:。 pos:助動詞 ne:O DP:nil
16 O NXT:CLS:X NXT:EOS pos:記号-句点 ne:O DP:nil
```

図 4.2: テストデータ抽出した存在物の１つずつに注目し，その存在物ごとに，対応する場所を検出するタスクとする．の例（次のクラスが X の場合）

手順 6： 文末から文頭の順に推定を行う.

推定の例を表 4.2 に示す．品詞は紙面の都合上，省略して記す.

存在物を抽出する場合を例に説明する．文末から推定を行うので，まず，単語「。」を推定する.「。」の品詞は記号であるため SVM は存在物でないと判断し,「。」のタグを「O」と推定する．次に,「。」の文頭側の単語「た」を推定する.「た」の次の推定タグの部分には,「た」の文末側の単語「。」の推定結果「O」が付与される.「た」の品詞は助動詞であり次の単語は「EOS（End Of Sentense)」であるため存在物でないと判断し,「た」のタグを「O」と推定する．次は「た」の文末側「まし」を推定する．このような順で SVM は推定を繰り返す.

表 4.2: タグの推定の例（存在物の抽出）

単語	素性					推定結果
	次の単語の推定タグ	次の単語	品詞	固有表現タグ	係り先	
名古屋	O	駅	名詞	B-LOCATION	nil	O
駅	O	で	名詞	I-LOCATION	nil	O
で	I	N	助詞	O	で撮影	O
N	I	7	記号	O	nil	I
7	I	0	名詞	O	nil	I
0	I	0	名詞	O	nil	I
0	E	系	名詞	O	nil	I
系	O	と	名詞	O	nil	E
と	I	ドクター	助詞	O	とドクター	O
ドクター	E	イエロー	名詞	O	nil	I
イエロー	O	を	名詞	O	nil	E
を	O	撮影	助詞	O	を撮影	O
撮影	O	し	名詞	O	nil	O
し	O	まし	動詞	O	nil	O
まし	O	た	助動詞	O	nil	O
た	O	。	助動詞	O	nil	O
。	X	EOS	記号	O	nil	O

手順 7： IOE2 タグを IOB2 タグに変換する．タグを B または I と推定したものが存在物/場所を抽出した箇所となる．表 4.2 の例では「Ｎ７００系」と「ドクターイエロー」が存在物として抽出されている.

4.3 実験

4.3.1 実験条件

第 3 章のコーパスを用いて実験を行う．提案手法は 8 分割のクロスバリテーションとする

4.3.2 実験結果—抽出性能

表 4.1 に抽出性能を評価した結果を示す．ここで，適合率 $P = pp/(pp + pn)$，再現率 $R = pp/(pp + np)$，F 値 $= 2PR/(P + R)$ である．また，pp は,「正解タグ B または I を，B または I と推定した数」，pn は,「正解タグ O を，B または I と推定した数」，np は,「正解タグ B または I を，O と推定した数」である．

表 4.3: 抽出実験の評価（単語単位）

手法	P	R	F 値	pp	pn	np
B_1（存在物）	0.49	0.02	0.03	32	33	1,518
提案（存在物）	0.84	0.06	0.11	94	17	1,456
B_1（場所）	0.84	0.60	0.70	530	96	348
提案（場所）	0.83	0.60	0.70	534	103	344

4.3.3 実験結果—チャンク単位

表 4.2 にチャンク単位の結果を示す．ここで，一致数は「チャンク内の正解タグ B または I 全てを，B または I と推定したチャンク数」である．

一致，および，不一致の例を以下に示す．表 4.1 の「N/7/0/0/系」の部分の正解タグは「B/I/I/I/I」である．例えば，推定結果が「B/I/I/I/I」となっていると一致となる．推定結果が「O/B/I/I/I」のように，B または I と推定したい箇所に一つでも O のタグを推定すると不一致となる．

表 4.4: 抽出実験の評価値（チャンク単位）

手法	一致率	一致数	不一致数
B_1（存在物）	0.03	15	551
提案（存在物）	0.03	17	549
B_1（場所）	0.64	294	164
提案（場所）	0.65	296	162

4.3.4 存在物の抽出結果

実際に抽出した存在物の例を図 4.3 に示す.

ドクターイエロー，キヤ検３２２，のぞみ，０系，７００系，こだま 特急きりしま，イーストアイ，メーテル（銅像），桃太郎（車両の名前） 黄色い新幹線，ワイドビュー伊那路

図 4.3: 抽出した存在物の例

数は少ないが，品詞が「名詞-一般」,「名詞-数」,「名詞-固有名詞-一般」などの一般名詞である単語を存在物として抽出したことを確認できた.

4.3.5 場所の抽出結果

実際に抽出した場所の例を図 4.4 に示す

新神戸，岡山駅，新大阪，名古屋，飯田線，豊橋，静岡県浜松市 コンコース，石山坂本線，ＪＲ生野駅，栄生駅，山口，門司機関区 鉄道・リニア館，青函トンネル，名神高速道，富士山剣ヶ峰

図 4.4: 抽出した存場所の例

地名や駅名など，固有表現タグに LOCATION があるものは抽出できた．他にも「コンコース」や「鉄道・リニア館」など一般名詞の抽出も確認できた.

チャンク単位で抽出できなかったものは，大きく分けて以下の 3 種類であった.

記号：「東京〜博多」の「〜」
助詞：「名古屋の駅」の「の」
一般名詞：「新富士駅付近」の「付近」

この結果から，チャンク内に固有表現タグ LOCATION がある単語がある場合，その他の単語が抽出されないことがわかった.

第5章　存在物と場所の対応検出

記事内にある存在物と場所の対応を SVM を用いて検出する．抽出した存在物の１つずつに注目し，その存在物ごとに，対応する場所を検出するタスクとする．

5.1　手法

5.1.1　ベースライン手法

ベースライン B_2，および，B_3 の 2 種類設ける．

B_2：注目する存在物から記事の先頭側と末尾側に向けて各単語を調べ，単語数による距離で最短の所にある場所の表現（B タグの語）を対応する場所とする．

B_3：全てのリンク先を対応する場所とする．

図 5.1 の記事を例に説明する．下線 E は存在物を，下線 L は場所を示す．まず B_2 の検出の説明する．存在物 E に着目する．文頭側にある場所は $L1$ で，E と $L1$ の単語距離は 24 である．文末側にある場所は $L2$ と $L3$ である．$L2$ の方が E との単語区間の距離が短いので，$L2$ に着目する．E と $L2$ の単語距離は 15 である．文末側の E と $L1$ の単語距離間の方が短いので,「E は $L1$ に存在する」と検出する．

B_3 は全てのリンク先を対応するので,「E は $L1$，$L2$，および，$L3$ に存在する」と検出する．

尼崎$_{L1}$/に/移動/。/
今度/は/道/に/迷い/ませ/ん/でし/た/。/
本当/なら/空が/青い/の/です/が/。/
上り/ドクター/イエロー$_{E}$/が/来る/１/分/前/に/飛ん/で/いき/まし/た/。/
この/あと/梅田/キャノン$_{L2}$/に/行き/大阪/駅$_{L3}$/を/ウロ/ウロ/と/。/

図 5.1: 記事の例（単語境界付き）

5.1.2 提案手法

1つの記事内全ての各場所を，注目する存在物とペアにして，各ペアが対応するべきか否かを，SVM で判定する．次の素性を用いる．

単語距離

f1 存在物と場所の単語距離が全ペアのうち最短か否か.

存在物/場所の品詞

f2 存在物と場所の品詞

動詞のペア

f3 存在物/場所の表現（チャンク）の係り先の動詞の基本形のペア.

f4 f3 に場所の表現が存在物から文頭側にあるか文末側にあるか表記したもの.

助詞

f5 場所の表現の直後の助詞.

存在物と場所の間にある単語

f6 存在物と場所の間にある動詞，さらに末尾側の存在物/場所から文末側にある動詞.

f7 存在物と場所の間にある動詞および助詞，さらに末尾側の存在物/場所から文末側にある動詞または文末になるまでたどって得られる動詞および助詞.

f8 存在物と場所の間にある名詞以外の単語，さらに末尾側の存在物/場所から文末側にある動詞または文末になるまでたどって得られる名詞以外の単語.

f9 存在物と場所の間にある単語，さらに末尾側の存在物/場所から文末側にある動詞または文末になるまでたどって得られる単語.

f10 存在物と場所の間にある動詞，助詞および名詞，さらに末尾側の存在物/場所から文末側にある動詞または文末になるまでたどって得られる動詞，助詞および名詞．ただし名詞は「名詞-サ変接続」と「名詞-副詞可能」に限る.

表現数

f11 存在物存在物と場所の間にある場所の表現数.

意味コード

f12 存在物や場所の表現を含む文に出現する動詞の意味コード（日本語語彙大系の一般名詞意味属性および用言意味属性）のペア.

f13 存在物や場所の表現を含む文に出現する名詞および動詞の意味コード（日本語語彙大系の一般名詞意味属性および用言意味属性）のペア.

提案手法は素性の組み合わせ方によりまずは次の 16 通りとする.

存在物/場所の品詞

M_1: f1 および f2 を用いる手法

動詞のペア

M_2: f1，f3 を用いる手法

M_3: f1，f4 を用いる手法

品詞+動詞のペア

M_4: f1，f2，および，f3 を用いる手法

品詞+助詞

M_5: f1，f2，および，f5 を用いる手法

品詞+存在物と場所の間にある単語

M_6: f1，f2，および，f6 を用いる手法

M_7: f1，f2，および，f7 を用いる手法

M_8: f1，f2，および，f8 を用いる手法

M_9: f1，f2，および，f9 を用いる手法

M_{10}: f1，f2，および，f10 を用いる手法

提案手法 M_7+表現数

M_{11}: f1，f2，f7，および，f11 を用いる手法

動詞のペア+助詞

M_{12}: f1，f3，および，f5 を用いる手法

動詞のペア+存在物と場所の間にある単語

M_{13}: f1，f3，および，f8 を用いる手法

M_{14}: f1，f3，および，f9 を用いる手法

動詞のペア+意味コード

M_{15}: f1，f3，および，f12 を用いる手法

M_{16}: f1，f3，および，f13 を用いる手法

さらに SVM のスコアが正値かつ最大値のペアを推定結果とする方法 M_{sgx}，および，正値のペアをすべて推定結果とする方法 M_{plx} を設ける $(x = 1, 2, ..., 16)$.

5.2 実験の様子

5.2.1 手順

手順 1： コーパスを分割する．

コーパスを 8 分割し，そのうち 1 つをテストデータ，他を学習データとする．

手順 2： 存在物と記事内の場所をそれぞれペアにする．

記事の例を図 5.2 に示す．下線 E は存在物を，下線 L は場所を示す．例の場合，存在物と場所のペアは，$(E, L1)$，$(E, L2)$，および，$(E, L3)$ の 3 ペアとなる．

尼崎$_{L1}$ に移動。
今度は道に迷いませんでした。
本当なら空が青いのですが。
上り ドクターイエロー$_{E}$ が来る 1 分前に飛んでいきました。
このあと 梅田キャノン$_{L2}$ に行き 大阪駅$_{L3}$ をウロウロと。

図 5.2: 記事の例

手順 3： 4.1.2 節で述べた素性を作成する．

手順 4： 学習データを SVM で学習する．

手順 5： テストデータを SVM に通し，存在物と場所の各ペアが対応しているか判定する．

手順 6： M_{sgx} で評価する場合は存在物がドクターイエローかつ，SVM のスコアが正値かつ最大値のペアの場所を検出する．M_{plx} で評価する場合は存在物がドクターイエローかつ SVM のスコアが正値のペアの場所を検出する．

図 5.1 の例の M_2 における SVM のスコアは，$(E, L1) = 0.99$，$(E, L2) = 1.74$，$(E, L3) = -0.26$ であった．M_{sg2} の場合は $L2$ を検出する．M_{pl2} の場合は $L2$，および，$L1$ を検出する．

5.3 実験

5.3.1 実験条件

コーパスの IOB2 タグを参照して存在物と場所を定め，それらの対応検出のみを評価する．提案手法は 8 分割のクロスバリテーションとする．

5.3.2 評価方法

実験の評価はリンク単位の場合と名称単位の場合を算出する．評価は，F 値で行い，各手法の F 値を比較する．評価基準として広く用いられている適合率，再現率の値も算出する．

適合率 $P = \langle$ 一致数 $\rangle / \langle$ 推定数 $\rangle$，再現率 $R = \langle$ 一致数 $\rangle / \langle$ 得られるべき数 $\rangle$，F 値 $= 2PR/(P + R)$ である．

5.3.3 実験結果—リンク単位の場合

存在物と場所の得られるべきリンク数は 2,240 であり，この数についての評価結果を表 5.1 に示す.

表 5.1: 対応検出の評価（リンク単位）

検出数	素性の組み合わせ	手法	P	R	F 値	一致数	推定数
単数	単語間の距離が最短	B_2	0.75	0.19	0.30	422	566
	存在物/場所の品詞	M_{sg1}	0.76	0.19	0.31	428	566
	動詞のペア	M_{sg2}	0.72	0.18	0.29	407	566
		M_{sg3}	0.73	0.18	0.29	411	566
	存在物/場所の品詞+動詞のペア	M_{sg4}	0.68	0.17	0.27	384	565
	存在物/場所の品詞+助詞	M_{sg5}	0.72	0.18	0.29	410	566
	存在物/場所の品詞+	M_{sg6}	0.64	0.16	0.26	358	562
	存在物と場所の間にある単語	M_{sg7}	0.61	0.15	0.24	343	562
		M_{sg8}	0.61	0.15	0.24	341	560
		M_{sg9}	0.53	0.13	0.21	298	564
		M_{sg10}	0.59	0.13	0.22	330	556
	提案手法 M_7+表現数	M_{sg11}	0.60	0.15	0.24	334	562
	動詞のペア+助詞	M_{sg12}	0.72	0.18	0.29	409	566
	動詞のペア+	M_{sg13}	0.63	0.16	0.25	351	560
	存在物と場所の間にある単語	M_{sg14}	0.60	0.15	0.24	341	566
	動詞のペア+	M_{sg15}	0.69	0.17	0.28	387	564
	意味コード	M_{sg16}	0.64	0.16	0.25	356	560
複数	存在物/場所の品詞	M_{pl1}	0.51	0.37	0.43	835	1,639
	動詞のペア	M_{pl2}	0.61	0.25	0.37	564	926
		M_{pl3}	0.62	0.24	0.35	538	866
	存在物/場所の品詞+動詞のペア	M_{pl4}	0.55	0.41	0.47	910	1,660
	存在物/場所の品詞+助詞	M_{pl5}	0.53	0.31	0.39	698	1,308
	存在物/場所の品詞+	M_{pl6}	0.50	0.55	0.52	1,228	2,460
	存在物と場所の間にある単語	M_{pl7}	0.49	0.59	0.53	1,319	2,702
		M_{pl8}	0.50	0.53	0.51	1,190	2,393
		M_{pl9}	0.45	0.57	0.50	1,278	2,850
		M_{pl10}	0.49	0.53	0.51	1,312	2,668
	提案手法 M_7+表現数	M_{pl11}	0.48	0.56	0.52	1,261	2,602
	動詞のペア+助詞	M_{pl12}	0.57	0.28	0.38	634	1,106
	動詞のペア+	M_{pl13}	0.47	0.52	0.49	1,271	2,510
	存在物と場所の間にある単語	M_{pl14}	0.46	0.56	0.50	1,247	2,725
	動詞のペア+	M_{pl15}	0.50	0.37	0.43	829	1,648
	意味コード	M_{pl16}	0.54	0.48	0.50	1,083	2,015
	全てのペア	B_3	0.42	1.00	0.59	2,240	5,363

P, R，および，F 値の最も高い値を比較する．まずは，M_{sgx} の評価に着目する．P は M_{sg1} が 0.76 となる．R は B_2，および，M_{sg1} が 0.19 となる．F 値は B_2，M_{sg1}，および，M_{sg2} が 0.31 となる．次に，M_{plx} の評価に着目する．P は M_{pl3} が 0.62 となる．R は B_3 が 1.00 となる．F 値は B_3 が 0.59 となる．よって，リンク単位の評価は B_3 が最も良い.

5.3.4 実験結果—名称単位の場合

得られるべき場所の文字列の異なり数は 95 であり，この数についての評価結果を表 5.2 に示す.

表 5.2: 対応検出の評価（名称単位）

検出数	素性の組み合わせ	手法	P	R	F 値	一致数	推定数
単数	単語間の距離が最短	B_2	0.82	0.57	0.67	54	66
	存在物/場所の品詞	M_{sg1}	0.78	0.56	0.65	53	68
	動詞のペア	M_{sg2}	0.76	0.54	0.63	51	67
		M_{sg3}	0.77	0.54	0.63	51	66
	存在物/場所の品詞+動詞のペア	M_{sg4}	0.75	0.52	0.61	49	65
	存在物/場所の品詞+助詞	M_{sg5}	0.76	0.55	0.64	52	68
	存在物/場所の品詞+	M_{sg6}	0.79	0.48	0.60	46	58
	存在物と場所の間にある単語	M_{sg7}	0.75	0.46	0.57	44	58
		M_{sg8}	0.78	0.48	0.60	46	59
		M_{sg9}	0.71	0.42	0.53	40	56
		M_{sg10}	0.94	0.55	0.70	53	56
	提案手法 M_7+表現数	M_{sg11}	0.74	0.45	0.56	43	58
	動詞のペア+助詞	M_{sg12}	0.76	0.54	0.63	51	67
	動詞のペア+	M_{sg13}	0.83	0.05	0.10	5	6
	存在物と場所の間にある単語	M_{sg14}	0.71	0.05	0.10	5	7
	動詞のペア+	M_{sg15}	0.71	0.52	0.59	49	69
	意味コード	M_{sg16}	0.71	0.05	0.10	5	7
複数	存在物/場所の品詞	M_{pl1}	0.65	0.65	0.65	62	96
	動詞のペア	M_{pl2}	0.60	0.63	0.62	60	100
		M_{pl3}	0.67	0.64	0.66	61	91
	存在物/場所の品詞+動詞のペア	M_{pl4}	0.53	0.72	0.61	68	129
	存在物/場所の品詞+助詞	M_{pl5}	0.61	0.68	0.64	65	107
	存在物/場所の品詞+	M_{pl6}	0.51	0.77	0.62	73	142
	存在物と場所の間にある単語	M_{pl7}	0.49	0.81	0.60	77	158
		M_{pl8}	0.53	0.77	0.62	73	138
		M_{pl9}	0.40	0.73	0.52	70	173
		M_{pl10}	0.53	0.83	0.65	79	148
	提案手法 M_7+表現数	M_{pl11}	0.48	0.78	0.60	74	154
	動詞のペア+助詞	M_{pl12}	0.59	0.69	0.63	66	112
	動詞のペア+	M_{pl13}	0.46	0.06	0.11	6	13
	存在物と場所の間にある単語	M_{pl14}	0.38	0.08	0.13	8	21
	動詞のペア+	M_{pl15}	0.46	0.65	0.54	62	136
	意味コード	M_{pl16}	0.82	0.13	0.22	12	15
	全てのペア	B_3	0.36	1.00	0.53	95	264

P, R，および，F 値の最も高い値を比較する．まずは，M_{sgx} の評価に着目する．P は M_{sg10} が 0.94 となる．R は B_2 が 0.57 となる．F 値は M_{sg10} が 0.70 となる．次に，M_{plx} の評価に着目する．P は M_{pl3} が 0.67 となる．R は B_3 が 1.00 となる．F 値は M_{pl3} が 0.66 となる．M_{10} は M_{plx} の R が 2 番目に高い値である．よって，M_{10} で使用した素性が有用であると考える．

5.3.5 検出結果

検出頻度の上位 10 件を表 5.3 に示す．上位 10 件のうち 9 件はドクターイエローが存在する場所である．

表 5.3: 対応検出の評価（名称単位）

順位	検出した場所	検出回数	正解
1	豊橋駅	10	○
1	東京駅	10	○
3	名古屋	9	○
4	富士川	8	○
4	西淀区	8	×
6	博多駅	7	○
6	中里	7	○
6	小倉駅	7	○
9	東京	6	○
9	新幹線ホーム	6	○

5.3.6 Google 検索との比較

一般的的に存在性情報を得るときは，本やインターネットを使用する．そこで，本研究では比較対象として Google 検索を用いる．

Google 検索で「ドクターイエロー 場所」と検索した．検索後に１ページ目にあるリンク先を１つずつ見ていった結果，ドクターイエローは東京駅〜博多駅の区間で見ることができることが分かった．つまり，東京，品川，新横浜，小田原，熱海，三島，新富士，静岡，掛川，浜松，豊橋，三河安城，名古屋，岐阜羽島，米原，京都，新大阪，新神戸，西明石，姫路，相生，岡山，新倉敷，福山，新尾道，三原，東広島，広島，新岩国，徳山，新山口，厚狭，新下関，小倉，および，博多の各駅で見ることができる．

提案手法では，駅名以外にも中里や富士川などを検出した．駅以外のドクターイエローが見えやすいスポットである．また検出頻度を出すことで東京駅〜博多駅の中では東京駅や豊橋駅が通過駅ではなく写真を撮ることができる場所であることが考えられ，Google 検索で出された候補を絞りこむこともできる．

よって提案手法が有用であることが確認できた．

第6章　異なるコーパスにおける対応検出

第5章ではドクターイエローコーパスで実験を行った．ドクターイエローは特殊な話題であるので，一般性を確認するため「お土産ブログ」のコーパスでも解析を行う．

6.1　コーパス

2013年4月のブログからお土産に関係する記事は112記事抽出された．文数2,943は，単語数51,349はとなり，存在物834ヶ所あり，場所は1,415ヶ所であった．存在物リンクの付与された場所は567ヶ所であった．存在物と場所のリンク数は15,012であった．

112記事のうち7記事は記事内に場所が1箇所しかなかった．リンク数にすると20であった．そのうち存在物が赤福のペアは3であり，2ペアは存在する場所のペアであった．

6.2　実験

6.2.1　実験条件

5.1.2提案手法で述べた16種類の手法を用いて，8分割クロスバリテーションで実験する．

6.2.2 実験結果—リンク単位の場合

存在物と場所の得られるべきリンク数は 2,496 であり，この数についての評価結果を表 6.1 に示す.

表 6.1: 対応検出の評価（リンク単位）

手法	P	R	F 値	一致数	推定数
B_2	0.56	0.19	0.28	468	834
M_{sg1}	0.55	0.18	0.28	459	834
M_{sg2}	0.53	0.18	0.27	444	832
M_{sg3}	0.51	0.17	0.26	425	832
M_{sg4}	0.53	0.18	0.26	439	832
M_{sg5}	0.54	0.18	0.27	451	832
M_{sg6}	0.47	0.16	0.23	387	824
M_{sg7}	0.50	0.16	0.25	410	827
M_{sg8}	0.47	0.15	0.23	382	818
M_{sg9}	0.48	0.16	0.24	390	814
M_{sg10}	0.48	0.16	0.24	392	824
M_{sg11}	0.55	0.18	0.28	458	828
M_{sg12}	0.54	0.18	0.27	449	833
M_{sg13}	0.46	0.15	0.23	377	821
M_{sg14}	0.46	0.15	0.23	379	828
M_{sg15}	0.50	0.17	0.25	412	828
M_{sg16}	0.45	0.15	0.22	366	821
M_{pl1}	0.53	0.19	0.28	464	874
M_{pl2}	0.48	0.20	0.28	494	1,020
M_{pl3}	0.48	0.20	0.28	503	1,055
M_{pl4}	0.48	0.19	0.28	485	1,021
M_{pl5}	0.54	0.18	0.27	456	842
M_{pl6}	0.31	0.29	0.30	731	2,392
M_{pl7}	0.30	0.31	0.31	770	2,552
M_{pl8}	0.26	0.33	0.29	818	3,091
M_{pl9}	0.27	0.34	0.30	842	3,149
M_{pl10}	0.31	0.30	0.30	754	2,466
M_{pl11}	0.36	0.35	0.36	869	2,391
M_{pl12}	0.49	0.20	0.28	492	1,010
M_{pl13}	0.25	0.31	0.28	771	3,061
M_{pl14}	0.27	0.31	0.29	782	2,896
M_{pl15}	0.45	0.20	0.28	504	1,119
M_{pl16}	0.32	0.22	0.26	554	1,727
B_3	0.23	1.00	0.38	2,496	15,012

P, R，および，F 値の最も高い値を比較する．まずは，M_{sgx} の評価に着目する．P は B_2 が 0.56 となる．R は B_2 が 0.19 となる．F 値は B_2，M_{sg1}，および，M_{sg11} が 0.28 となる．次に，M_{plx} の評価に着目する．P は M_{pl5} が 0.54 となる．R は B_3 が 1.00 となる．F 値は B_3 が 0.38 となる．よって，リンク単位の評価は B_3 が最も良い.

6.2.3 実験結果—名称単位の場合

存在物「赤福」について得られるべき場所の文字列の異なり数は 20 であり，この数についての評価結果を表 6.2 に示す.

表 6.2: 対応検出の評価（名称単位）

手法	P	R	F 値	一致数	推定数
B_2	0.69	0.55	0.61	11	16
M_{sg1}	0.67	0.50	0.57	10	15
M_{sg2}	0.63	0.50	0.56	10	16
M_{sg3}	0.65	0.55	0.59	11	17
M_{sg4}	0.63	0.50	0.56	10	16
M_{sg5}	0.71	0.50	0.59	10	14
M_{sg6}	0.73	0.40	0.51	8	11
M_{sg7}	0.73	0.40	0.51	8	11
M_{sg8}	0.67	0.40	0.50	8	12
M_{sg9}	0.77	0.50	0.61	10	13
M_{sg10}	0.75	0.45	0.56	9	12
M_{sg11}	0.75	0.45	0.56	9	12
M_{sg12}	0.69	0.55	0.61	11	16
M_{sg13}	1.00	0.05	0.10	1	1
M_{sg14}	1.00	0.05	0.10	1	1
M_{sg15}	0.69	0.55	0.61	11	16
M_{sg16}	0.73	0.40	0.51	8	11
M_{pl1}	0.69	0.55	0.61	11	16
M_{pl2}	0.67	0.60	0.63	12	18
M_{pl3}	0.67	0.60	0.63	12	18
M_{pl4}	0.65	0.55	0.59	11	17
M_{pl5}	0.73	0.55	0.63	11	15
M_{pl6}	0.43	0.50	0.47	10	23
M_{pl7}	0.43	0.50	0.47	10	23
M_{pl8}	0.69	0.55	0.61	11	16
M_{pl9}	0.65	0.55	0.59	11	17
M_{pl10}	0.62	0.50	0.55	10	16
M_{pl11}	0.48	0.60	0.53	12	25
M_{pl12}	0.71	0.60	0.65	12	17
M_{pl13}	1.00	0.05	0.10	1	1
M_{pl14}	1.00	0.15	0.26	3	3
M_{pl15}	0.69	0.55	0.61	11	16
M_{pl16}	0.69	0.55	0.61	11	16
B_3	0.23	1.00	0.38	20	84

P, R，および，F 値の最も高い値を比較する．まずは，M_{sgx} の評価に着目する．P は M_{sg13}，および，M_{sg14} が 1.00 となる．R は B_2，M_{sg3}，M_{sg12}，および，M_{sg15} が 0.55 となる．F 値は M_{sg12}，および，M_{sg15} が 0.61 となる．次に，M_{plx} の評価に着目する．P は M_{pl13}，および，M_{sg14} が 1.00 となる．R は B_3 が 1.00 となる．F 値は M_{pl12} が 0.65 となる．よって，M_{12} で使用した素性が有用であると考える．

6.2.4 検出結果

「赤福」に対応する場所を図 6.1 に示す.

伊勢，おはらい町，おかげ横丁，名古屋
大阪市北区梅田３－１－３ＪＲ大阪三越伊勢丹
神戸そごう，おかげ横町，ＪＲ大阪三越伊勢丹店，ＪＲ大阪三越伊勢丹，門前町
名阪上野ドライブイン，名古屋駅，東海地方，池袋西部，赤福本店，関西
伊勢神宮そばの通り，伊勢神宮，ひろしま菓子博，おはらい町おかげ横丁

図 6.1: 正しい検出例

表 6.2 より，F 値が最も高い M_{pl12} の検出結果を図 6.2 に示す.

伊勢，私鉄，おはらい町，門前町，名阪上野ドライブイン，名古屋駅
名古屋，東北，東海地方，地元の駅，神戸そごう，駅構内，伊勢神宮
デパ地下，ひろしま菓子博，おかげ横丁，ＪＲ大阪三越伊勢丹店

図 6.2: M_{pl12} の検出結果

6.2.5 Google 検索との比較

Google 検索で「赤福 場所」で検索すると，赤福の公式ホームページが見つかった．赤福公式ホームページには全店舗一覧があったのでそのページと比較する．赤福公式ホームページによると，三重県，愛知県，大阪府，京都府，兵庫県，奈良県，岐阜県，および，滋賀県では常に購入できる店舗があることがわかった.

しかし，提案手法では名阪上野ドライブイン，ひろしま菓子博，神戸そごうなどでも購入できることが新たにわかった.

第7章　オープンテスト

第5章では，ドクターイエローコーパスを用いてクロスバリテーションで実験を行った．第6章では，お土産コーパスを用いてクロスバリテーションで実験を行った．

クロスバリテーションでの実験は学習データがドクターイエローやお土産の存在する場所を学習している可能性がある．そこで，学習データをドクターイエローコーパス，テストデータをお土産コーパスにして作成したシステムが有用であるかを確認する．

7.1　実験

7.1.1　実験条件

5.1.2 提案手法で述べた 16 種類の手法で実験を行う．学習データにドクターイエローコーパス、テストデータにお土産コーパス用いる．

7.1.2 実験結果—リンク単位の場合

存在物と場所の得られるべきリンク数は 2,496 であり，この数についての評価結果を表 7.1 に示す.

表 7.1: 対応検出の評価（リンク単位）

手法	P	R	F 値	一致数	推定数
B_2	0.56	0.19	0.28	468	834
M_{sg1}	0.43	0.14	0.21	357	834
M_{sg2}	0.50	0.17	0.25	418	834
M_{sg3}	0.47	0.16	0.23	390	834
M_{sg4}	0.41	0.14	0.21	345	834
M_{sg5}	0.41	0.14	0.21	342	834
M_{sg6}	0.39	0.13	0.19	319	819
M_{sg7}	0.37	0.12	0.19	308	823
M_{sg8}	0.38	0.13	0.19	313	830
M_{sg9}	0.37	0.12	0.19	309	826
M_{sg10}	0.40	0.13	0.20	330	823
M_{sg11}	0.41	0.14	0.20	340	824
M_{sg12}	0.47	0.16	0.24	393	834
M_{sg13}	0.40	0.13	0.20	332	831
M_{sg14}	0.38	0.13	0.19	317	831
M_{sg15}	0.47	0.16	0.23	388	833
M_{sg16}	0.45	0.15	0.22	369	828
M_{pl1}	0.22	0.35	0.28	884	3,932
M_{pl2}	0.35	0.22	0.27	555	1,594
M_{pl3}	0.33	0.30	0.27	574	1,749
M_{pl4}	0.22	0.35	0.27	874	4,037
M_{pl5}	0.20	0.38	0.26	949	4,789
M_{pl6}	0.20	0.42	0.27	1,045	5,278
M_{pl7}	0.18	0.54	0.26	1,342	7,663
M_{pl8}	0.17	0.55	0.26	1,365	7,806
M_{pl9}	0.18	0.57	0.28	1,429	7,865
M_{pl10}	0.21	0.55	0.31	1370	6407
M_{pl11}	0.19	0.55	0.28	1,377	7,306
M_{pl12}	0.25	0.27	0.26	682	2,781
M_{pl13}	0.17	0.59	0.26	1,468	8,851
M_{pl14}	0.18	0.57	0.28	1,429	7,714
M_{pl15}	0.25	0.33	0.28	832	3,375
M_{pl16}	0.23	0.40	0.29	989	4,311
B_3	0.23	1.00	0.38	2,496	15,012

P, R，および，F 値の最も高い値を比較する．まずは，M_{sgx} の評価に着目する．P は B_2 が 0.56 となる．R は B_2 が 0.19 となる．F 値は B_2 が 0.28 となる．次に，M_{plx} の評価に着目する．P は M_{pl2} が 0.35 となる．R は B_3 が 1.00 となる．F 値は B_3 が 0.38 となる．よって，リンク単位の評価は B_3 が最も良い.

7.1.3 実験結果—名称単位の場合

存在物「赤福」について得られるべき場所の文字列の異なり数は 20 であり，この数についての評価結果を表 7.2 に示す.

表 7.2: 対応検出の評価（名称単位）

手法	P	R	F 値	一致数	推定数
B_2	0.69	0.55	0.61	11	16
M_{sg1}	0.44	0.35	0.39	7	16
M_{sg2}	0.69	0.55	0.61	11	16
M_{sg3}	0.75	0.60	0.67	12	16
M_{sg4}	0.53	0.40	0.46	8	15
M_{sg5}	0.56	0.45	0.50	9	16
M_{sg6}	0.53	0.40	0.46	8	15
M_{sg7}	0.50	0.40	0.44	8	16
M_{sg8}	0.60	0.45	0.51	9	15
M_{sg9}	0.58	0.35	0.44	7	12
M_{sg10}	0.60	0.45	0.51	9	15
M_{sg11}	0.47	0.35	0.40	7	15
M_{sg12}	0.73	0.55	0.62	11	15
M_{sg13}	1.00	0.05	0.09	1	1
M_{sg14}	1.00	0.05	0.09	1	1
M_{sg15}	0.73	0.55	0.62	11	15
M_{sg16}	0.58	0.35	0.44	7	12
M_{pl1}	0.37	0.55	0.44	11	30
M_{pl2}	0.70	0.60	0.65	12	17
M_{pl3}	0.70	0.60	0.65	12	17
M_{pl4}	0.37	0.55	0.44	11	30
M_{pl5}	0.32	0.60	0.42	12	37
M_{pl6}	0.57	0.60	0.58	12	21
M_{pl7}	0.46	0.65	0.54	13	28
M_{pl8}	0.44	0.75	0.56	15	34
M_{pl9}	0.44	0.60	0.51	12	27
M_{pl10}	0.46	0.65	0.54	13	28
M_{pl11}	0.43	0.65	0.52	13	30
M_{pl12}	0.52	0.60	0.55	12	23
M_{pl13}	1.00	0.15	0.26	3	3
M_{pl14}	0.75	0.15	0.25	3	4
M_{pl15}	0.30	0.65	0.41	13	43
M_{pl16}	0.29	0.50	0.36	10	35
B_3	0.23	1.00	0.38	20	84

P, R，および，F 値の最も高い値を比較する．まずは，M_{sgx} の評価に着目する．P は M_{sg13}，および，M_{sg14} が 1.00 となる．R は M_{sg3} が 0.60 となる．F 値は M_{sg3} が 0.67 となる．次に，M_{plx} の評価に着目する．P は M_{pl13} が 1.00 となる．R は B_3 が 1.00 となる．F 値は M_{pl2}，および，M_{pl3} が 0.65 となる．よって，M_3 で使用した素性が有用であると考える．

第8章　考察

8.1　存在物の抽出と場所の抽出

8.1.1　存在物の抽出

ドクターイエローコーパスは未知語や，N700A などの英数字のみの表現は推定できなかった．これは単語境界を取得した時に「N/7/0/0/A」と分けられることが原因と考えられる．

8.1.2　場所の抽出

「東京〜博多」のように LOCATION タグを与えられる単語が含まれるチャンクは,「東京」と「博多」とに分かれて抽出される.「東京〜博多」と,「東京」と「博多」の抽出では意味が変わってしまうので,「〜」の部分も抽出することは必要である．一度抽出した結果を素性に加え，もう一度 SVM に推定させ,「東京〜博多」を抽出を行う必要がある．

8.2　存在物と場所の対応検出

8.2.1　素性

素性 f1 の「存在物と場所の単語距離が全ペアのうち最短か否か.」の影響が強い．最短のものが対応していない場所でも，SVM の結果では対応するとなることが多い．また素性の組み合わせや，書きかたを変更し，最適な素性を見つけなければならない．

8.2.2　存在する時間

赤福の場合，存在する場所に行けばいつでも購入できる（ただし売り切れや休業日などの特殊な場合を除く）．しかし，ドクターイエローはいつ行っても見ることができるとは限らない．ドクターイエローだけでなくイルミネーションなど観光支援の場合,「どこ

に何があるのか」という情報に加え,「いつ存在しているのか」という情報も重要になってくる．今後の課題として，存在物と場所の対応に時間情報を検出することも必要である．

8.2.3　場所から存在物を検出

今回は「ドクターイエローはどこで見ることができるのか」,「赤福はどこで買うことができるのか」のように存在物から存在する場所を検出することを行った．

しかし，場所から存在物を検出することは行っていない．旅行の計画を立てる場合，行き先が決まっていて「名物は何か」調べることがある．このように，場所から存在物を検出することも観光支援のために必要である．

第9章　おわりに

本研究では，観光支援として存在性情報の抽出を行った．抽出方法はSVMを用いて文章から存在物と場所の抽出，および，それらの対応を検出を行った．

存在物の抽出では，固有表現タグが付与されていない一般名詞の抽出を行った．実験を行った結果，固有表現タグより多く場所と存在物が抽出できた．

存在物と場所の検出では，存在物と記事内にある場所をそれぞれペアにし，各ペアに対して存在物と場所が対応しているかをSVMで識別した．

ドクターイエローコーパスを用いてクロスバリテーション，お土産コーパスを用いてクロスバリテーション，および，学習データにドクターイエローコーパスを，テストデータにお土産コーパスを用いる3種類の実験条件で行った．リンク単位での検出結果はベースラインのF値が0.59で一番高くなった．しかし，表現単位での検出結果は提案手法(M_3)のF値が0.66となった．

Google検索の結果と提案手法の比較を行った．ドクターイエローコーパスの場合，Google検索で得ることができた存在する場所は駅名がほとんどであった．しかし，提案手法では駅名の他にも，富士川や中里などの存在する場所も得ることができた．

お土産コーパスの実験でF値の向上を確認できたこと，Google検索との比較でGoogle検索で得られない場所を得られたことから，提案手法に対する一定の評価を得ることができたと考える．今後の課題は，場所から存在物の対応検出を行うこと，および，時間の存在する時間（いつ見ることができるか）の情報抽出を行うことである．

謝辞

徳久雅人講師には，終始に渡り研究の進め方や本論文の書き方など，細部に渡る御指導を頂きました．ここに深く感謝いたします．

また，本研究を進めるに当たり，種々の御助言を頂きました村田真樹教授，および，村上仁一准教授に心から御礼申し上げます．

その他様々な場面で御助力をいただいた計算機工学 C 講座の学生皆様に感謝の意を表します．

参考文献

[1] 北尾祐樹: “２文からの場所と存在物の解析”, 鳥取大学工学部知能情報工学科卒業論文, 2013.

[2] 笹野遼平, 黒橋禎夫: “Japanese Named Entity Recognition Using Non-local Information”, 情報処理学会論文誌, Vol.29, No.11, pp.3765-3776, 2008.

[3] TinySVM: Support Vector Machines. http://chasen.org/taku/software/TinySVM/

[4] 工藤拓, 松本裕治: “Support Vector Machines　を用いた chunk 同定”, 自然言語処理, Vol.9, No.5, pp.3-22, 2002.

[5] 池原悟, 宮崎正弘, 白井諭, 横尾昭男, 中岩浩巳, 小倉健太郎, 大山芳史, 林良彦: “日本語語彙大系”, 岩波書店, 1997.

[6] 白井諭, 大山芳史, 池原悟, 宮崎正弘, 横尾昭男: “日本語語彙大系について”, 情報処理学会研究報告, Vol.98, No.106, pp.47-52, 1998.

[7] CaboCha: Yet Another Japanese Dependency Structure Analyzer. https://code.google.com/p/cabocha/

[8] IREX 実行委員会（編）: “IREX ワークショップ予稿集”, 1999.