文章からの存在物と存在場所の抽出

菊池 春香 † 徳久 雅人 † 村田 真樹 † 村上 仁一 †
鳥取大学工学部 知能情報工学科 †

1 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．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．そこで本稿では，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することを目的とする．

2 コーパスの作成

2.1 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．

手順 2 記事内の文を CaboCha で構文解析し，単語，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 IOB タグを人手で付ける．

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 は記事単位でユニークとする．

2.2 コーパスの例

例文「名古屋駅で N700 系とドクターイエローを撮影しました」に注釈付けした例を表 1 に示す．

表 1 注釈付けの例

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	

2.3 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．

3 存在物と場所の抽出

3.1 手法

3.1.1 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．

3.1.2 提案手法

SVM を用いて文末から文頭の順に各単語の IOB 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 IOB 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．

3.2 実験

第 2 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．

表 2 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．

Extraction of existing objects and their locations from sentences

†Department of Information and Knowledge Engineering, Faculty of Engineering, Tottori University

表 2　抽出実験の評価

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　存在物と場所の対応検出

4.1　手法

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．

4.1.1　ベースライン手法

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_2)．

4.1.2　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．

f1　存在物と場所の単語距離が全ペアのうち最短か否か．

f2　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．

f3　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．

f4　場所の表現の直後の助詞．

f5　存在物と場所の間にある単語と，各ペアの末尾側の存在物/場所から文末側にある動詞または文末までの単語．

提案手法は素性の組み合わせ方によりまずは次の 3 通りとする．

M_1：f1 および f2 を用いる手法

M_2：f1，f2，および，f3 を用いる手法

M_3：f1，f2，および，f4 を用いる手法

M_4：f1，f2，および，f5 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 ($x = 1, 2, 3, 4$)．

4.2　実験

4.2.1　実験条件

コーパスの IOB 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．

4.2.2　実験の様子

下線 E は存在物を，下線 L は場所を示す．$L1$ だけが正しい対応先である．$L2$ は E から最短である．M_2 における SVM のスコアは，$L1 = 0.99$，$L2 = 1.74$，$L3 = -0.26$ であった．B_2 と M_{sg2} は $L2$ を選択し（誤り），M_{pl2} は $L1$ と $L2$ を選択した（一部分誤り）．

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。

4.2.3　実験結果

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 3 に示す．ドクターイエローの存在場所について，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 4 に示す．ここで，適合率 P = ⟨ 一致数 ⟩/⟨ 推定数 ⟩，再現率 R = ⟨ 一致数 ⟩/⟨ 得られるべき数 ⟩ である．

前者の結果より，F 値では M_{pl2} や M_{pl4} が優れるが，後者の結果によるとその限りではない．特に M_{pl2} では「加島」という特定の表現が目立った．機械学習により特定の語が集められたためと考える．

表 3　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	0.75	0.19	0.30	422	566
M_{sg1}	0.72	0.18	0.29	407	566
M_{sg2}	0.64	0.16	0.25	356	560
M_{sg3}	0.72	0.18	0.29	409	566
M_{sg4}	0.60	0.15	0.24	341	566
M_{pl1}	0.61	0.25	0.37	564	926
M_{pl2}	0.54	0.48	0.50	1,083	2,015
M_{pl3}	0.57	0.28	0.38	634	1,106
M_{pl4}	0.46	0.56	0.50	1,247	2,725

表 4　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	0.82	0.57	0.67	54	66
M_{pl1}	0.60	0.63	0.62	60	100
M_{pl2}	0.82	0.13	0.22	12	15
M_{pl3}	0.59	0.69	0.63	66	112
M_{pl4}	0.38	0.08	0.13	8	21

5　おわりに

本稿では，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する方法を提案した．実験において，提案手法の再現率の向上が見られた．

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.