地名表現の使われ方における多義性の解消

小島 正裕† 村田 真樹‡ 西村 涼† 渡辺 靖彦†
†龍谷大学 理工学部 情報メディア学科
t060570@mail.ryukoku.ac.jp, r_nishimura@afc.ryukoku.ac.jp, watanabe@rins.ryukoku.ac.jp
‡独立行政法人 情報通信研究機構
murata@nict.go.jp

1 はじめに

日本語で書かれた文書を英語に機械翻訳する場合や、質問応答システムから知識を取り出す場合の問題として、単語の多義性の問題がある。単語の多義性の問題は、古くから自然言語処理における重要な問題の 1 つとして位置付けられており、これまでに様々な多義性を解消する試みが報告されている [1, 2, 3] 。以下に示すように日本語の地名表現は、場所的意味や組織的意味といった多義性がある。

(例文 1a) 夏の 日本 は小笠原気団におおわれて蒸し暑い日が続く。

(例文 1b) 3 大会連続出場の 日本 の戦いが注目される。

(例文 1a) では、その場所の気候を解説している文になり、この場合の「日本」は場所的意味になる。一方、(例文 1b) では、その組織の状況を解説している文になり、「日本」は (例文 1a) と同じ表記であるが、この場合の「日本」は場所的意味ではなく、組織的意味になる。このような地名表現において、地名表現の多義性を正しく解釈しておかなければ、機械翻訳システムで正しく翻訳されない。例えば、(例文 1a) の「日本」は「Japan」と翻訳し、(例文 1b) の「日本」は「Japan Team」と翻訳すべきである。このように同じ「日本」という表記であっても、文によって意味が異なる場合がある。

また、単語の多義性の問題は機械翻訳システムだけでなく、質問応答システムにおいても重要である。ユーザの質問で用いられている地名表現に対して、多義性を解消しておかなければ、質問応答システムから、適切な知識を得られない場合がある。例えば、以下のような質問文から回答文を取り出した場合、

質問文 日本 はワールドカップに出ますか。

回答文 日本 で開催されたサッカーのワールドカップで、…

質問文で用いられている「日本」は、チームを表しており、この質問文においての「日本」は組織的意味である。このような質問文に対して、「日本」を場所的意味で用いている回答文からは、適切な知識が得られない。

このような問題に対して、われわれは、機械学習を用いて、地名表現における多義性の解消を行う。また機械学習として、SVM と MEM を用いる。

2 実験データ

2.1 前処理

実験データは、京大コーパス [4] の 1995 年 1 月 1 日から 1995 年 1 月 10 日までの記事データに含まれる、日本語で用いられている地名表現 5,461 個を用いた。地名表現 5,461 個に対して、その地名表現が場所的意味で用いられているか、組織的意味で用いられているかについてタグ付けを行った。実験データを毎日新聞の記事を時系列順に上から下に並べ、データの下半分をデータ A とし、データの上半分をデータ B とする。表 1 に、データ A、B に、場所的意味で用いられた地名表現と組織的意味で用いられた地名表現の内訳を示す。

表 1: 地名表現における意味の割合

	データ A		データ B	
場所的意味	89.93%	(2,455/2,730)	89.20%	(2,436/2731)
組織的意味	9.71%	(265/2,730)	10.22%	(279/2,731)
その他	0.36%	(10/2,730)	0.58%	(16/2731)

2.2 場所的意味と組織的意味

実験データについて調査を行ったところ、以下のように整理できることがわかった。

大陸 (例) アジア

(例文 2a) アジア におけるサッカーの運営・管理・普及活動を行う。

(例文 2b) アジア が欧州破り初優勝

(例文 2a) の「アジア」という地名表現は場所的意味で用いている。これに対して、(例文 2b) の「アジア」という地名表現は、アジア地域を代表する選手で構成されるチームの意味で用いている。このため、(例文 2b) での「アジア」は、組織的意味で用いている地名表現といえる。

国 (例) 日本

(例文 3a) 開催条件として、現在の 日本 にはない 8 万人以上収容の競技場が必要と文書に明記した。

(例文 3b) 10 年前に 日本 はアジア通貨基金 (AMF・Asian Monetary Fund) の設立を主張していた。

(例文 3a) の「日本」という地名表現は、場所的意味で用いている。これに対して、(例文 3b) の「日本」という地名表現は、日本を管理している日本政府の意味で用いている。このため、(例文 3b) での「日本」は、組織的意味で用いている地名表現といえる。

都道府県・市町村 (例) 和歌山市

(例文 4a) 三日午後一時九分ごろ、和歌山市 で震度 3 の地震があった。

(例文 4b) 和歌山市 は、二年後から投棄を開始する。

(例文 4) のように地名表現には、大陸・国だけではなく、都市もある。(例文 4a) の「和歌山市」は、和歌山市という場所で地震があったので、場所的意味で用いている。これに対して、(例文 4b) の「和歌山市」は、和歌山市の自治体を意味している。このため、(例文 4b) での「和歌山市」は、組織的意味で用いている地名表現といえる。

地域 (例) 湘南

(例文 5a) 私、湘南 に移住します。

(例文 5b) 中田氏が古巣・湘南 の練習に参加。

(例文 5) で用いられている「湘南」のように、行政区分と一致しない地域名でも組織的意味で用いる場合がある。(例文 5a) は、「湘南」という場所に移住することを表しているので、場所的意味で用いている。これに対して、(例文 5b) は、湘南というチームについて表している。このため、(例文 5b) での「湘南」は、組織的意味で用いている地名表現といえる。

施設 (例) 東京拘置所

(例文 6a) 東京拘置所 で一夜を明かすことになります。

(例文 6b) 東京拘置所 が開示を拒否したため保全されなかった。

人によって作られた施設も地名表現の対象となる。(例文 6a) は、「東京拘置所」という場所で一夜を明かしたことを表しているので、「東京拘置所」は場所的意味で用いている。これに対して、(例文 6b) は、東京拘置所という組織を表している。このため、(例文 6b) での「東京拘置所」は、組織的意味で用いている地名表現といえる。

このように、地名表現が、国や大陸といった場所そのものを指していたり、人が常駐していないような場合は、場所的意味で用いられ、地名表現が、代表・管理・運営する人または団体を表す意味で用いられている場合は、組織的意味で用いられる場合が多い。しかし、人が常駐していないような地名表現においても、まれに組織的意味で用いることがある。例えば、以下の (例文 7) は、「金星」について述べている文である。

例外 (例) 金星

(例文 7a) 先に到達できるのは 金星 です。

(例文 7b) 金星 応答なし

(例文 7a) では、「金星」という場所をさしており、人が常駐しておらず場所的意味で用いている。しかし、(例文 7b) の「金星」は、人が常駐していないが場所的意味ではなく、「金星」という組織からの応答がないと表しているといえ、組織的意味で用いている。

以上のような、地名表現における多義性の解消を行う。

3 実験で用いる素性

機械学習によって多義性の解消を行うために用いる素性を表 2 に示す。この素性は多義性を解消したい地名表現を持つ本文から取り出す。なお、分類番号とは、分類語彙表 [5] に記されている語の意味ごとに与えられる 10 桁の番号のことである。単語の意味が近いものに対してこの 10 桁の番号も近くなる。単語の中でも意味が複数あるものにはその数だけ番号がふられている。実験では番号を 5 桁、3 桁、に区切り素性として与えている。これにより、それぞれの単語の上位概念を素性とすることができる [6]。

素性をグループ分けしたものを表 3 に示す。グループはさらに 4 つに大別できる。

G1、G2、G3 は、多義性の解消を行う対象となる地名表現を含む文節に関する情報を利用している。多義性の解消を行うためには、地名表現が含まれている文において、地名表現の前後にある情報が有効だと考えられるからである。

G4、G5、G6 と G7、G8、G9 は、文の構造を明らかにするため、多義性の解消を行う対象となる地名表現における係り元や係り先の情報を利用している。G4、G5、G6 は、「ワールドカップに参加する日本は …」という例文の場合、「参加する」という用言によって、組織的な意味として使われる地名表現であると学習させることができる。これは、係り元の情報が有効に働く例である。また G7、G8、G9 は、「日本で開催されるワールドカップに …」という例文の場合、「開催される」という用言によって、場所的な意味で使われている地名表現であると学習させることができる。これは、係り先の情報が有効に働く例である。なお、係り受け解析器として、cabocha [7] を用いる。

G10 は、多義性の解析を行う対象となる地名表現の、前後にある文字列の情報を利用している。

4 実験と検討

4.1 ベースライン手法

提案手法の比較手法として、ベースライン手法を用いる。ベースライン手法は、以下の手順で場所的意味か組織的意味かについて判定を行う。2.2 節で用意したデータ A、データ B を利用した。

Step1 データ A に判定の対象となる地名表現がある場合
データ B において、

場所的意味としての使用頻度の方が多い場合

場所的意味と判定し、処理を終了する。

組織的意味としての使用頻度の方が多い場合

組織的意味と判定し、処理を終了する。

場所的意味と組織的意味の使用頻度が同じ場合

データ B で多く使用する意味の方に判定し、

処理を終了する。

Step2 データ A に判定の対象となる地名表現がない場合
データ B で多く使用する意味の方に判定し、処理を終了する。

Step3 Step1 と Step2 で判定できなかった場合
実験データにあるすべての地名表現で、最も多く使用する意味に判定し、処理を終了する。

以上の処理をデータ A の 2,730 個に対して行い、提案手法の評価を行う。

表 2: 多義性の解消に用いる素性

素性	説明	素性	説明
s1	多義性の解消を行う地名表現を含む文節の最初の自立語		を含む文節に係る文節に含まれる付属語
s2	s1 の品詞	s38	s37 の品詞
s3	s1 の自立語の分類語彙表での分類番号 5 桁までの数字	s39	多義性の解消を行う地名表現を含む文節に係る文節に含まれる記号
s4	s1 の自立語の分類語彙表での分類番号 3 桁までの数字	s40	s39 の品詞
s5	多義性の解消を行う地名表現を含む文節の最後の自立語	s41	多義性の解消を行う地名表現を含む文節が係る文節の最初の自立語
s6	s5 の品詞		
s7	s5 の自立語の分類語彙表での分類番号 5 桁までの数字	s42	s41 の品詞
s8	s5 の自立語の分類語彙表での分類番号 3 桁までの数字	s43	s41 の自立語の分類語彙表での分類番号 5 桁までの数字
s9	多義性の解消を行う地名表現を含む文節の自立語	s44	s41 の自立語の分類語彙表での分類番号 3 桁までの数字
s10	s9 の品詞		
s11	s9 の自立語の分類語彙表での分類番号 5 桁までの数字	s45	多義性の解消を行う地名表現を含む文節が係る文節の最後の自立語
s12	s9 の自立語の分類語彙表での分類番号 3 桁までの数字	s46	s45 の品詞
s13	多義性の解消を行う地名表現を含む文節の最初の付属語	s47	s45 の自立語の分類語彙表での分類番号 5 桁までの数字
s14	s13 の品詞		
s15	多義性の解消を行う地名表現を含む文節の最後の付属語	s48	s45 の自立語の分類語彙表での分類番号 3 桁までの数字
s16	s15 の品詞		
s17	多義性の解消を行う地名表現を含む文節の付属語	s49	多義性の解消を行う地名表現を含む文節が係る文節の自立語
s18	s17 の品詞		
s19	多義性の解消を行う地名表現を含む文節の記号	s50	s49 の品詞
s20	s19 の品詞	s51	s49 の自立語の分類語彙表での分類番号 5 桁までの数字
s21	多義性の解消を行う地名表現を含む文節に係る文節の最初の自立語	s52	s49 の自立語の分類語彙表での分類番号 3 桁までの数字
s22	s21 の品詞		
s23	s21 の自立語の分類語彙表での分類番号 5 桁までの数字	s53	多義性の解消を行う地名表現を含む文節が係る文節の最初の付属語
s24	s21 の自立語の分類語彙表での分類番号 3 桁までの数字	s54	s53 の品詞
s25	多義性の解消を行う地名表現を含む文節に係る文節の最後の自立語	s55	多義性の解消を行う地名表現を含む文節が係る文節の最後の付属語
s26	s25 の品詞	s56	s55 の品詞
s27	s25 の自立語の分類語彙表での分類番号 5 桁までの数字	s57	多義性の解消を行う地名表現を含む文節が係る文節の付属語
s28	s25 の自立語の分類語彙表での分類番号 3 桁までの数字	s58	s57 の品詞
s29	多義性の解消を行う地名表現を含む文節に係る文節の自立語	s59	多義性の解消を行う地名表現を含む文節が係る文節の記号
s30	s29 の品詞	s60	s59 の品詞
s31	s29 の自立語の分類語彙表での分類番号 5 桁までの数字	s61	多義性の解消を行う地名表現の前 1,2,3,4,5 文字
s32	s29 の自立語の分類語彙表での分類番号 3 桁までの数字	s62	多義性の解消を行う地名表現の先頭から後方に向かって取り出した 1,2,3,4,5 文字
s33	多義性の解消を行う地名表現を含む文節に係る文節の最初の付属語	s63	多義性の解消を行う地名表現の後尾から前方に向かって取り出した 1,2,3,4,5 文字
s34	s33 の品詞		
s35	多義性の解消を行う地名表現を含む文節に係る文節の最後の付属語	s64	多義性の解消を行う地名表現の後 1,2,3,4,5 文字
s36	s35 の品詞		
s37	多義性の解消を行う地名表現	s65	多義性の解消を行う地名表現

4.2 地名表現の多義性の解消の実験

本節では、4.1 節で述べたベースライン手法と、提案手法と、素性選択について実験を行った結果について述べる。提案手法による機械学習と素性選択の実験は、データ A をテストデータ、データ B を学習データとして行い、正解率と F 値を求めた。素性選択の実験には、村田らが採用している手法 [8] を用い、データ B において 10 分割クロスバリデーションを行い、取り除く素性を決めた。また、SVM には TinySVM [1] の線形カーネルを利用し、ソフトマージンパラメータを 1 とした。また、MEM には maxent [2] を用いた。素性選択により取り除かれた素性は、SVM の場合は S4、S5、S9、S11、S33、S57、S65 であり、MEM の場合は S19、S47、S57 であった。それぞれの機械学習で共通して取り除かれた素性は、S57 の素性であった。表 4 に、種々の手法におけるデータ A に対する実験結果を示す。この表のベースライン手法の F 値から、地名表現が、場所的意味かどうかの判定をすることは、あまり難しくないが、組織的意味かどうかの判定をすることは、難しいことがわかる。機械学習による手法は、SVM、MEM ともにベースライン手法よりも正解率、F 値ともに高く、組織的意味の判定においても、良好な F 値であり、提案手法が有効であることがわかる。また、素性選択で求めた素性を実験結果では、正解率、F 値が SVM、MEM すべてにおいて高くなることはなかった。この結果から、素性選択で求めた素性は、データ A に対する分類にあまり役立たなかったと考えられる。このため、全素性を利用するのがよいことがわかる。

符号検定を用いて、ベースライン手法を用いた結果と機械学習を用いた結果に対して、統計的に有意差があるかどうかを調べた。まず、オープンデータに対して、ベースライン手法を適用した結果と、素性選択で求めた素性を用いた SVM の結果をそれぞれ比較した。ベースライン手法では不正解であり、SVM では正解であった地名表現は 136 個あった。一方、SVM では不正解であり、ベースライン手法では正解であった地名表現は 51 個であった。次に、オープンデータに対して、ベースライン手法を適用した結果と、全素性を用いた MEM の結果をそれぞれ比較した。ベースライン手法では不正解であり、MEM では

表 3: グループ化した素性

グループ	素性	説明
G1	s1-s12	対象となる地名表現を含む文節にある自立語に関する情報
G2	s13-s18	対象となる地名表現を含む文節にある付属語に関する情報
G3	s19-s20	対象となる地名表現を含む文節にある記号に関する情報
G4	s21-s32	対象となる地名表現を含む文節に係る文節の自立語に関する情報
G5	s33-s38	対象となる地名表現を含む文節に係る文節の付属語に関する情報
G6	s39-s40	対象となる地名表現を含む文節に係る文節の記号に関する情報
G7	s41-s52	対象となる地名表現を含む文節が係る文節の自立語に関する情報
G8	s53-s58	対象となる地名表現を含む文節が係る文節の付属語に関する情報
G9	s59-s60	対象となる地名表現を含む文節が係る文節の記号に関する情報
G10	s61-s65	対象となる地名表現とその前後の文字列

表 4: 種々の手法における実験結果

手法	利用した素性	正解率	F 値	
			場所的意味	組織的意味
ベースライン	全素性	90.51%	0.949	0.188
SVM	全素性	93.22%	0.963	0.619
	素性選択	93.30%	0.963	0.612
MEM	全素性	93.68%	0.965	0.613
	素性選択	93.59%	0.965	0.610

[1] http://chasen.org/~taku/software/TinySVM/

[2] http://www2.nict.go.jp/x/x161/members/mutiyama/software.html#maxent

正解であった地名表現は 141 個あった。一方、MEM では不正解であり、ベースライン手法では正解であった地名表現は 65 個あった。これらの結果に対して、比率の差の検定を行ったところ、両方とも有意水準 5%(両側検定) で有意差があった。

また、ベースラインの手法による実験で、地名表現の意味の判定が不正解だったが、機械学習の手法による実験で正解とすることができた例について検討を行った。

正解例 昨年十月一日未明、武蔵野市のスナックで倒れ、病院に収容された。

この例文の「武蔵野市」という地名表現は都市のことを意味するので場所的意味となる。ベースラインの手法では、この地名表現を組織的意味として判定した。これは、データ A、B において、場所的意味で用いられる割合よりも組織的意味で用いられる割合が多かったからである。使われている意味の頻度の情報だけで判定するのではなく、判定したい地名表現を含む文の前後の情報を利用して、判定を行う方が良いことを示唆している。

4.3 実験で用いた素性の検討

提案手法に用いた素性のグループにおける有意差の分析を行った。全グループの素性を用いた場合の出力と、全グループの素性から 1 つのグループの素性を省いた場合の出力を比較し、ブートストラップ法 (反復数 10,000) を用いてそれらの有意差を調べた。データ B のみを利用した 10 分割クロスバリデーションによる検討実験 1、データ A をテストデータ、データ B を学習データによる検討実験 2 を行った。検討実験 1、2 には、全素性を用いた場合に正解率がよかった MEM を用いた。

表 5 に全グループの素性を用いた場合の F 値が高かった回数 (勝利回数) と、全グループの素性から 1 つのグループを省いた場合の勝利回数を示す。検討実験 1、2 ともに、全グループの素性を用いた場合の勝利回数が 9,500 回を超えるグループは、G1 においての組織的意味を判定する場合であった。G1 は組織的意味を判定する場合に特に有効なグループであることがわかり、このグループを用いることで、組織意的意味の F 値を 0.034(0.569 → 0.613) 向上させることがわかった。

また、組織的意味の判定において、G1 の素性を追加することで不正解から、正解とすることができた例について検討を行なった。

正解例 …、米国の入国審査もメキシコ側のチェックも厳しくなった。

この例文の地名表現である「メキシコ」が含まれる文節の自立語は、' メキシコ'、' 側' である。MEM で求まる α 値を正規化した値 [9] を求めたところ、組織的意味を判定する場合、上位に G1 の素性が多くあった。's9:側' の情報が高い値を示し、この例文のように、地名表現が含まれる文節に' 側' があると、地名表現が' その組織においての' といった意味になる場合が多かった。このように、G1 の素性には、地名表現が組織的意味である場合に、使われる情報が多いため、有効なのではないかと考えられる。

5 おわりに

地名表現の多義性の問題を解決するために、機械学習 (SVM、MEM) を用いた手法を提案した。

表 5: ブートストラップ法による素性の分析

省いたグループ	全グループ			
	検討実験 1		検討実験 2	
	場所的意味	組織的意味	場所的意味	組織的意味
G1	323	10,000	402	9,994
G2	1,135	7,773	4,482	2,342
G3	1,869	6,337	2,466	6,028
G4	8,009	1,437	4,516	5,528
G5	988	6,062	7,675	3,438
G6	1,818	6,034	4,177	0
G7	9,466	7,729	8,142	2,963
G8	2,835	568	1,693	5,511
G9	3,973	4,016	1,476	6,034
G10	947	6,988	985	5,419
省いたグループ	1 種類			
	検討実験 1		検討実験 2	
	場所的意味	組織的意味	場所的意味	組織的意味
G1	9,461	0	9,331	3
G2	7,996	1,413	4,436	6,683
G3	6,028	0	5,863	187
G4	1,136	7,964	4,638	3,546
G5	7,727	1,817	992	3,534
G6	6,048	1,871	4,185	0
G7	331	1,668	1,368	6,271
G8	574	307	7,524	2,428
G9	611	1,843	7,326	1,764
G10	338	2,283	93	3,806

提案手法を新聞記事に対して実験した結果、SVM より MEM の方が良い結果であった。MEM において正解率 93.68%、場所的意味の F 値 0.965、組織的意味の F 値 0.613 が得られた。また、ベースラインの手法との比較において、正解率、F 値ともに良い結果が得られた。特に、組織的意味で用いられている地名表現の分類においては、ベースラインの手法よりも F 値で 0.425 高い値が得られた。また、符号検定を行ったところ、本報告で提案した手法で得られた結果は、ベースラインの手法で得られた結果よりも、有意水準 5%(両側検定) で優れていることを確認した。

素性のグループに対して検討を行ったところ、MEM において、組織的意味では G1 のグループが有効であることを確認し、F 値を 0.034 向上させた。

参考文献

[1] 池原 悟, 村上 仁一, 桐澤 洋: “意味的用法に着目した日本語名詞の英訳語選択について”, 情報処理学会論文誌, Vol.44, No.5, pp. 1343–1353, (2003).

[2] 村田 真樹, 内山 将夫, 内元 清貴, 馬 青, 井佐原 均: “SENSEVAL2J 辞書タスクでの CRL の取り組み”, 自然言語処理, Vol.9, No.2, (2002).

[3] 佐藤充, 森辰則: “事実型質問応答における画像・地図を用いた回答提示”, 言語処理学会第 13 回年次大会発表論文集, pp. 752–755, (2007).

[4] 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, pp. 115–118, (1997).

[5] 国立国語研究所 編: “分類語彙表”, 大日本図書株式会社, (2004).

[6] 村田 真樹, 神埼 享子, 内元 清貴, 馬 青, 井佐原 均: “意味ソート msort”, 自然言語処理, Vol.2, No.3, (1995).

[7] 工藤 拓, 松本 裕治: “チャンキングの段階適用による係り受け解析”, 情報処理学会論文誌, Vol. 43, No. 6, pp. 1834–1842, (2002).

[8] 村田 真樹, 金丸 敏幸, 白土 保, 井佐原 均: “入力文の格助詞ごとに学習データを分割した機械学習による受動態の能動態への変換における格助詞の変換”, システム制御情報学会論文誌, Vol.21, No.6, pp. 165-175, (2008).

[9] Masaki Murata, Ryo Nishimura, Kouichi Doi, Toshiyuki Kanamaru and Kentaro Torisawa: “Analysis of the Degree of Importance of Information Using Newspapers and Questionnaires”, IEEE NLPKE-08, pp.122-139, (2008).