ブログから観光開発案を分析する際の分析者間の比較

高原明日美 *1 徳久雅人 *2 村上仁一 *2 村田真樹 *2
*1 鳥取大学 工学部 知能情報工学科
*2 鳥取大学大学院 工学研究科 情報エレクトロニクス専攻
*1,*2{s112029, tokuhisa, murakami, murata}@ike.tottori-u.ac.jp

1 はじめに

観光地の開発案を考えるために，旅行について書かれたブログ記事を参考にすることが挙げられる．ブログ記事は，日々多くの人々に書かれているため，特定の観光地に絞ったとしても，すべてに目を通すことはできない．そこで，ブログ記事の中から開発案につながる文 (ヒント文) を自動抽出し，開発案の発想支援が行われている．例えば，ルールに基づいてヒント文を抽出する手法 [1]，SVM や能動学習を用いる手法 [2]，[3] がある．

ここで，ヒント文の判定は分析者の主観に強く依存する．一般的に主観的分析における分析者間の一致数の評価は κ 値が使われているが，ヒント文の判定では文脈の幅で差が生じるという問題がある．

そこで，本稿では，分析者間での判定結果を比較する方法について検証することを目的とする．

2 ヒント文の分析の概要

ヒント文の分析では，分析者がブログ文を読み，今後の観光地開発のヒントになるかどうかを判定する．

コーパスとして糸魚川ブログデータを用いる (表 1)．このデータは Yahoo!ブログの「旅行」の項目で「糸魚川 観光」という検索キーで得られた 95 記事，3,222 文である．検索は 2011 年 10 月 19 日に行われた．

分析者はこのコーパスを見てヒント文か非ヒント文かを判定 (+1/−1 を文に付与) する．ヒント文と判定した場合，ヒントカテゴリを 1 つ選ぶ．このヒントカテゴリは [1] で使用されていた 17 分類である．

先行研究 [2]，[3] では，分析者 1 人で本コーパスを対象としてヒント文を抽出していたが，本稿では，分析者を 3 人に増やした．その結果を表 2 にまとめる．a 者は 1595 文，b 者は 812 文，c 者は 275 文をヒント文として判定した．ヒントカテゴリ毎に付与件数および合計数の差が見られた．本稿では，こうした分析者間の差が妥当であるかどうかを検討する．

表 1: ブログデータの例

id	文
I000000	確か２年ぶりの晴山ゴルフ場。
I000001	会社関係で２０人弱でのコンペ。
I000002	初めてコースを周る初心者が数名いるので、このコースは距離が短いので最適かな。
I000003	東京から距離も近いし地方から転勤で来た人も観光地の軽井沢を案内できるので、何かと便利。
I000004	乗用カートは無いので、手引きカート。

表 2: ヒントカテゴリ毎の付与件数

	ヒントカテゴリ	分析者		
		a 者	b 者	c 者
1	自然散策	170	138	39
2	動植物	39	7	5
3	飲食	251	104	60
4	買い物	54	63	3
5	街並み	49	31	1
6	施設	192	101	19
7	温泉	169	85	51
8	神社仏閣	16	22	10
9	文化歴史	107	55	36
10	音楽	4	2	0
11	スポーツ・アウトドア	23	17	5
12	釣り	9	1	2
13	交流	13	6	0
14	産業	3	2	0
15	交通	347	124	30
16	行事	46	41	13
17	その他	103	13	1
	合計	1595	812	275

3 文脈を考慮しないコーパスの比較

3.1 κ 値による比較

3 人のうち任意のペアについて文ごとにヒント文か非ヒント文かを比較して一致数を抽出し，κ 値 [4] を求める．

その結果，3 人それぞれがヒント文として抽出したものにはかなりばらつきが見られた (表 3)．そして κ 値は 3 つともかなり低い値になり，κ 値を基準にするならばヒント付けの結果はほぼ一致していないと言える．

表 3: ヒント文か非ヒント文かにおける κ 値

	a 者対 b 者	a 者対 c 者	b 者対 c 者
Po	0.642	0.567	0.784
Pe	0.502	0.504	0.706
κ	0.280	0.126	0.267

3.2 ヒントカテゴリに注目した一致数

ヒントカテゴリの一致を集計した (表 4)．3 者ともがヒント文だとした文について集計する (168 文が該当)．

表 4: 分析者 3 名がヒントありとしたもの

ヒントカテゴリの一致/不一致	件数
3 人とも不一致	10
2 人だけ一致	45
3 人とも一致	113
合計	168

表 5: 3 人のヒントカテゴリが不一致の例文

ヒントカテゴリ			文
a 者	b 者	c 者	
施設	街並み	文化歴史	「志摩」や「懐華楼」など，江戸時代そのままに残されたお茶屋建物もあり見学できます
街並み	飲食	施設	太閤山ランドと海王丸パークという富山県射水市の自然豊かな素晴らしいスポットを案内してもらい，夕方小杉駅近くのショッピングセンターでお茶をしてから，近くの寿司屋にいって新鮮な刺身や寿司に舌鼓を打っている間に...
動植物	施設	行事	8 月 21 日までは夏季特別展なので，「イルカショー」も開催♪

表 4 より，同じ文でも分析者によりヒントカテゴリが異なる結果になったものは 55 件 (10+45 件) あった．しかし，表 5 を見ると，1 つの文に複数のカテゴリの要素が含まれていると読み取れるので，カテゴリの不一致が生じたと解釈できる．

3.3 z 得点による比較

z 得点とは平均が 0，標準偏差が 1 になるようにカテゴリ毎の文数を変換して得られる得点である．各者で独立して算出する．

図 1 にヒントカテゴリ毎の文数の z 得点を示す．横軸はカテゴリ番号，縦軸は z 得点である．図 1 を見ると，カテゴリ 10〜14 では 3 者ともが同じような比率になっている．a 者と c 者は同じような比率になっているものが多いと読みとれる．a 者はカテゴリ 15 の交通，b 者はカテゴリ 1 の自然散策，c 者はカテゴリ 3 の飲食，カテゴリ 7 の温泉が他の分析者よりも多い比率で観光開発に役立つと判断したとわかる．

また，グラフは相似であると判断した．カテゴリ 4 と 17 を除き，3 者とも平均以上または平均以下となっているからである．このことから分析者 3 人が読みとった内容は類似しているといえる．

文単位で評価した κ 値では 3 者の差は大きかったが，z 得点により，度数が正規化されたことで相似であるという結果となった．

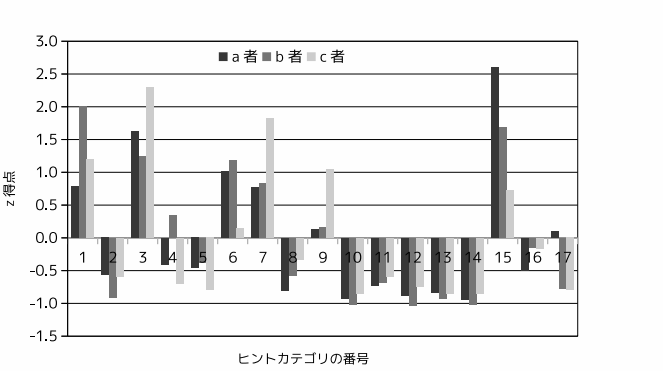

図 1: ヒントカテゴリ毎の文数の z 得点の比較

4 文脈を考慮した比較

一致性の評価には一般的に κ 値が使われている．しかし文脈がある中で κ 値を使用すると，本質的には良いアノテーションであっても，あまり良い評価は得られなかった．そこで，κ 値以外の方法で，文脈中での一致性を求める．

4.1 島の定義

分析者が連続してヒント文とした文の並びを「島」と呼び，並んだヒント文の数を「島の大きさ」と呼ぶことにする．島を単位とした一致性を比較する．比較対象の分析者の島に対して基準者が少なくとも 1 文でも同意していると，その島は「同意」が得られたとみなす．

表 6 に島の例を示す．ヒントありは+1，ヒントなしは −1 である．a 者は id001〜004，006〜008 を連続しヒントありとしているので，id001〜004，id006〜008 それぞれが a 者の島である．b 者は id002〜004，c 者は id003〜004 をそれぞれヒントありとしており，それぞれが島となる．a 者の島の大きさは 4 と 3 である．表 7 に a 者の島の例を示す．

表 6: 島の例

id	ヒント		
	a	b	c
001	+1	−1	−1
002	+1	+1	−1
003	+1	+1	+1
004	+1	+1	+1
005	−1	−1	−1
006	+1	−1	−1
007	+1	−1	−1
008	+1	−1	−1
009	−1	−1	−1

表 7: a 者の島の具体例

id	+1/−1	ヒント	文
I000413	−1		旅人はどんな思いでここを走ったのでしょうね。
I000414	+1	文化歴史	明治１６年（1883）に開通した国道８号線は、断崖の中腹を切り開いて造られました。
I000415	+1	文化歴史	今は「親不知コミュニティロード」と名付けられ、土木遺産、日本の道百選に選ばれています。
I000416	+1	文化歴史	一枚岩に彫られた「如砥如矢」の文字。
I000417	+1	文化歴史	砥石のように滑らかで矢のように真っ直ぐであるという意味だそうですが、厳しい絶壁を切り開いて、国道が完成した喜びを刻んだとのこと。
I000418	+1	文化歴史	この工事に尽力した青海の人、富岳磯平の書だそうです。
I000419	+1	文化歴史	ウォルター・ウェストンが訪れたことを機縁して像が立っていました。
I000420	+1	街並み	青海八景「四世代道暮色」と名付けられ名勝です。
I000421	+1	街並み	ここから見る風景は、初代「旧北陸道」、二世代目「天嶮開削道」、三世代目「天嶮トンネル」、四世代目「北陸自動車道」が同時に見られます。
I000422	+1	交通	高速道は、陸地に造る余地がまったくなく、海上にせり出して造られています。
I000423	+1	施設	展望台から見る情景は、まさに暮色に包まれていました。
I000424	−1		この日も暮れようとしています。

4.2 島の数と大きさ

分析者それぞれの島の大きさを求め，ヒストグラムを作成する．図 2 にその結果を示す．横軸は島の大きさ，縦軸は度数 (島の数) である．

図 2 より，最も大きな島での文数は a 者では 23 文，b 者では 18 文，c 者では 6 文であった．c 者と比べると，a 者と b 者は 1 つの島の大きさが 10 文以上のものが多い．

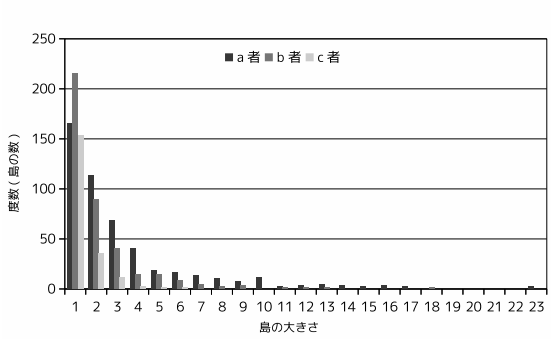

図 2: 島のヒストグラム

4.3 島の同意率

評価対象者を X 者，基準者を Y 者とする．X 者の島に対する Y 者の同意率 $A_{X,Y}$ を次式で求める．

$$A_{X,Y} = \frac{Y \text{ 者が同意している } X \text{ 者の島の数}}{X \text{ 者の島の数}}$$

この方法で X 者 Y 者に対して a 者 b 者 c 者を全通り適用する．また，a 者に対して「b 者または c 者」，b 者に対して「a 者または c 者」，c 者に対して「a 者または b 者」も，同じように求める．

4.4 結果

表 8 に 1 人対 1 人の同意率を示す．表 9 に基準者を 2 人とした同意率を示す．ここで，(i) は基準者が同意した対象者の島の数，(ii) は対象者の島の数である．

表 8: 2 者の場合の文脈中での一致性

対象者	基準者	(i)	(ii)	同意率 (i)/(ii)
a 者	b 者	277	482	0.57
a 者	c 者	140	482	0.29
b 者	c 者	147	393	0.37
b 者	a 者	334	393	0.85
c 者	a 者	180	204	0.88
c 者	b 者	160	204	0.78

表 9: 3 者の場合の文脈中での一致性

対象者	基準者	(i)	(ii)	同意率 (i)/(ii)
a 者	b 者または c 者	293	482	0.61
b 者	a 者または c 者	351	393	0.89
c 者	a 者または b 者	199	204	0.98

表 8 より，a 者がヒント付けした箇所 (島) で b 者が同意したのは，半分しかない．また，a 者に対して c 者は 3 割しか同意していない．もともと a 者は広い範囲 (全体の 50%) にヒント付けしているのだが，b 者 c 者に対して a 者が同意した割合が 8 割以上と高いのはある程度の一致性があったと評価できる．

表 9 より，a 者は約 6 割，b 者は約 9 割，c 者はほぼ全部が他の分析者から同意を得ている．c 者は少ないヒント付けにも関わらず，ほぼ全て同意されているので，必要最低限だけヒント付けしたと評価する．また，a 者は 4 割が不同意であり，無駄なヒント付けがあったと評価する．

5　考察

5.1　島の大きさによる分析者への制限付け

ヒントを多く付ける場合，ヒント判定は文脈から得られる情報を使用しており，観光開発案がわかりにくいことがある．
1 章で述べた通り, そもそも本コーパスは，ヒント文抽出のトレーニングデータにしておきたかった．そこで，ヒントの付けすぎを減らすために，連続する島の文数を c 者の最大文数の 6 文に制限することが，タグ付けが安定するための一案として考えられる．

ここで，表 10 に c 者の島の例を示す．c 者は a 者と比べると比較的島が小さい．idI003017 の文で「遺産」とあるが，その具体例が I003018 と I003019 に書かれている．c 者のように，具体例だけ注目すれば，開発内容が考えられるので，3 文に+1 を付けるよりも，2 文に+1 を付ける方がよいという考え方もある．

表 10: c 者の島の具体例

id	+1/−1	ヒント	文
I003017	−1		日本に、そして世界に誇れる地質遺産の数々…
I003018	+1	自然散策	明星山の大岩壁は約３億年前のサンゴ礁が変化したもので、多くの化石を含んでいます。
I003019	+1	自然散策	大岩壁の下では、小滝川の清流に洗われたヒスイが観察できます。

※ b 者は上記 3 文に+1 自然散策を付与.

5.2　自動抽出の性能と課題

分析者 3 人分のデータを実験データとして，ヒント文抽出の性能を確認する．手法 [1] では，3 人のデータで多数決をとり，3 人中 2 人以上がヒント文であると判断した文をヒント文としているので，本稿も同様の方法で自動抽出を行なう．

抽出方法には能動学習を用いる．まず，コーパスの先頭 5%を初期学習データとして抽出し，分析者に見せる (以下の%は分母を最初のコーパスのサイズとする)．残りの 95%は SVM でスコアが高いものから順に並び替える．その 95%の中から上位 5%を抽出して分析者に見せる．今，分析者は 10%を閲覧したことになる．また残りの 90%を SVM でスコアが高いものから順に並び替える．これを閲読率が 30%になるまで繰り返す．

その結果を表 11 に示す．ここで，適合率 P = (2 者以上の一致文数)/(ヒント文として自動抽出された文数)，再現率 R = (2 者以上の一致文数)/(正解データでのヒント文数)，F 値= $2PR/(P+R)$ である．分析者 1 人と 3 人のどちらともでは，ルールベース抽出より能動学習による抽出の方が全ての値が上昇しているので，能動学習により抽出性能が向上していることが確認できる．

本稿の結果を応用するとすれば，自動抽出結果における島 (推定された島) の大きさを 6 以下に制限するという方法が考えられる．つまり，島の中でスコアが高くても，島の中で上位 6 位までしか抽出しないことにする．この実現は今後の課題とする．

表 11: 自動抽出の結果

分析者数	手法	P	R	F 値	閲読率
1 人	ルール	0.70(638/911)	0.40(638/1595)	0.51	28.2%
1 人	能動学習	0.76(734/965)	0.46(733/1595)	0.57	30.0%
3 人	ルール	0.35(318/908)	0.44(318/723)	0.39	28.2%
3 人	能動学習	0.43(419/975)	0.58(419/723)	0.49	30.0%

6　おわりに

本稿では，主観的分析の一つである観光開発案のヒント文の判定について，分析者間の比較を行う方法を示した．文脈依存性のある判定であったので，次の点に注目した．

- z 得点による度数の正規化
- 「島」による文の並びの取り出し

z 得点により，ヒント文とした文数の絶対数に依存することなく，分析内容（ヒントカテゴリ）の偏りを分析者間で比較できた．島により，文脈に依存して分析した範囲をコーパスから抽出することが可能になり，島を文脈の近似的な単位として，分析者間の一致性（同意率）が評価できた．効果として，z 得点により内容の分布の類似性が認められ，かつ，同意率により一致性が認められる場合には，アノテーションの絶対数が少ない分析者に対しても妥当性があると評価することができた．また，応用として，分析の際や，自動抽出の際に，文脈に強く依存することを抑制するために島の大きさを制限することを挙げた．

参考文献

[1] 徳久雅人, 奥村秀人, 村田真樹: 観光開発のためのブログ記事からの評判分析, 観光と情報, Vol.7, No.1, pp.85–98, 2011.

[2] 徳久雅人, 村田真樹: 観光開発のヒントをブログ記事から得るための支援技術～SVM を用いる場合～, 第 8 回観光情報学会全国大会発表概要集, pp.44–45, 2011.

[3] 謝花博, 徳久雅人, 村田真樹, 村上仁一: 観光開発のヒントをブログ記事から得るための支援技術～能動学習を用いる場合～, 言語処理学会第 18 回年次大会発表論文集, pp.1324–1327, 2012.

[4] Cohen, J.: “A coefficient of agreement for nominal scales”, *Education and Psychological Measurement*, Vol.20, No.1, pp.37-46, 1960.