研究指導　中澤 真 教授

観光スポット推薦システムのための印象語抽出手法

西郷 藍

1. はじめに

旅行に行く際，旅行会社のパンフレットを用いて観光スポットの情報を集める代わりに，インターネットを使って情報を集めるようになった．JTB総合研究所の調査[1]によると，旅行前にスマートフォンで旅行先の下調べをする人の割合が高いことが示されている．ネットを使って観光情報の下調べをする場合，地名やランドマーク名をキーワードとした検索をするのが典型的である．一方，具体的な目的地を決めずにどこかに出かけたい場合は，ユーザが目的地として求める「涼しい」や「綺麗」などの印象を表す言葉で検索し，該当する場所についての情報が得られるのが望ましい．しかし，一般的な検索エンジンにおいて印象語による検索をしても，観光スポットのまとめサイトが結果として表示されるのみで，ユーザがイメージした言葉に適した結果が必ずしも提示されるわけではない．

印象語による情報検索や推薦システムを実現するために鍵となるのは，検索・推薦の結果となる観光スポットと印象語をどのように対応付けるかということである．観光スポットの公式サイトの文書内に，必ずしも旅行者が抱く印象語を含んでいるわけではないため，全文検索機能だけで解決するわけではないからである．このため，各観光スポットに対応する印象語をいかに抽出するかという問題に取り組むことが重要である．

印象語を抽出する研究はすでに行われており[2]，ブログを情報源として，記事に含まれる名詞，形容詞，動詞を取り出すことにより印象語を抽出している．しかし，名詞，形容詞，動詞の語すべてを印象語とすると，印象を表す語として適切ではない語が多く含まれてしまう．

そこで，本研究では印象語として用いる品詞を細分化し，印象語の出現率を考慮した品詞の並びパタンから，各観光スポットの印象語をより的確に抽出する手法を提案する．

2. 印象語による検索・推薦システムの現状

印象語による検索や推薦システムでは，ユーザが具体的な目的が決まっておらず漠然としたイメージしか持っていない場合でも情報を探すことができるのが特徴である．楽曲の検索や動画の推薦など様々な分野で印象語による情報検索・推薦システムの実現に向けた取り組みがなされている[3][4]．

実際，一般的な検索エンジンにおいて印象語で検索しても，ユーザがイメージした楽曲や動画が的確に得られるわけではない．これは観光分野においても同様で，ユーザが目的地として求める「涼しい」や「綺麗」などの印象語で検索しても，イメージ通りの観光スポットのサイトを検索結果として得られない．また，印象には個人差があるため，観光スポットの公式サイトに記載された文書内の印象語が，ユーザがイメージする印象語とは合致しない可能性がある．

この問題に対して，多数の人の印象が集まっているソーシャルメディア上のクチコミを情報源として印象語を抽出することで，多様な印象語の抽出と観光スポットの対応関係の明確化を図ることができると考えられる．

3. 印象語の抽出方法の課題

観光スポットについての印象語の抽出方法についてはいくつかの研究がある．伊達ら[2]はブログを情報源として，記事に含まれる名詞，形容詞，動詞を取り出すことにより印象語を抽出している．また山田[5]もブログを情報源としている点は同じであるが，語の並びのパタンに注目し，特定の品詞の並びの中に形容詞が現れたときのみ，その語を印象語として抽出している．

しかし，伊達らの研究では名詞，形容詞，動詞であれば無条件で印象語としているため，動詞の「進む」や「運ぶ」のように印象を表す語として適切ではない語が多く含まれる．また，名詞と形容詞においても細分化すると自立，接尾，非自立というように区分され，形容詞の接尾として「ぽい」，形容詞の非自立として「づらい」というように，単独では意味をなさない言葉も多く抽出してしまう．一方，山田の研究では品詞の並びのパタンによって印象語と判定される形容詞は限定されているが，品詞の並びのパタンを選ぶ際に印象語の出現率などを考慮していないため，適切なものを選択できていない可能性がある．

そこで，本研究では品詞を細分化したとき物事の性質や状態を表す形容詞（自立），名詞（形容動詞語幹）を印象語とし，印象語の出現率を考慮した品詞の並びのパタンから印象語を抽出する手法を提案する．

4. クチコミの分析と考察

先に述べたように，観光スポットの公式サイトでは紹介する側の印象に基づく記述であるため，旅行者の率直な感想や印象が反映されているとは限らず，観光スポットと印象語の対応関係の構築に適しているとはいえない．そのため，観光スポットに対する多数の人の感想や印象が書かれた旅行サイトなどのクチコミを用いることが，多様な印象語の抽出に適している．

4.1 クチコミの収集

本研究では国内の観光クチコミサイトである「じゃらん

net」[1]のクチコミを使用する．対象とするデータは2017年11月から2018年10月までの1年間に投稿されたものである．この期間内のクチコミから学習用データとして城郭ジャンルの鶴ヶ城，神社・神宮・寺院ジャンルの金閣寺，特殊地形ジャンルの秋芳洞の3ヶ所のクチコミ合計525件，検証用データとして動物園・植物園ジャンルの上野動物園，テーマパーク・レジャーランドジャンルのハウステンボス，山岳ジャンルの函館山の3ヶ所のクチコミ合計1359件を収集した[2]．検証用データのカテゴリをあえて別のものとしたのは，複数ジャンルの学習用クチコミデータに基づき印象語抽出の判定条件を作り，この条件が異なるジャンルの印象語抽出にも有用であることを示すためである．また，学習用データはクチコミ件数の多寡が影響しないように，1年分のクチコミの件数が同程度の観光スポットを選んだ．検証用データは観光スポットごとに評価できるように，1年分の件数が学習用データの総数と同程度のクチコミ件数である観光スポットを選んだ．

なお，クチコミの収集と分析には統計分析ソフト「R」[3]，形態素解析ソフト「MeCab」[4]を用いる[6][7]．

4.2 印象語とその傾向

印象語を抽出するにあたり，この言葉を厳密に定義する必要がある．観光の分野に限れば，観光スポットの印象や観光スポットで旅行者が感じた感情を表す言葉を印象語とするのが妥当である．そこで本研究では，形容詞（自立）と名詞（形容動詞語幹）のいずれかの品詞の語の中で，人が見たり聞いたりしている時に感じ取られたもの，または観光スポットで感じた感情のいずれかを表現した語を印象語と定義する．この定義に基づいて学習用データの各観光スポットから選んだ印象語を出現頻度順に表 1に示す．

表 1　選んだ印象語

観光スポット	印象語	品詞	頻度	印象語	品詞	頻度
鶴ヶ城	綺麗	名詞(形容動詞語幹)	29	立派	名詞(形容動詞語幹)	10
	良い	形容詞(自立)	27	暑い	形容詞(自立)	10
	美しい	形容詞(自立)	18	寒い	形容詞(自立)	10
	広い	形容詞(自立)	12	素敵	名詞(形容動詞語幹)	8
	素晴らしい	形容詞(自立)	11	楽しい	形容詞(自立)	8
金閣寺	美しい	形容詞(自立)	38	素敵	名詞(形容動詞語幹)	7
	綺麗	名詞(形容動詞語幹)	21	すごい	形容詞(自立)	6
	素晴らしい	形容詞(自立)	13	広い	形容詞(自立)	6
	良い	形容詞(自立)	12	豪華	名詞(形容動詞語幹)	4
	有名	名詞(形容動詞語幹)	9	悪い	形容詞(自立)	4
秋芳洞	涼しい	形容詞(自立)	39	暗い	形容詞(自立)	14
	広い	形容詞(自立)	20	良い	形容詞(自立)	13
	自然	名詞(形容動詞語幹)	18	寒い	形容詞(自立)	12
	いい	形容詞(自立)	16	大きい	形容詞(自立)	12
	暑い	形容詞(自立)	15	すごい	形容詞(自立)	12

印象語の定義では形容詞（自立）と名詞（形容動詞語幹）を条件としているが，この二つの品詞の単語すべてが印象語となるわけではない．これを確認するために印象語である語を1，印象語ではない語を0と手動でラベル付けし，出現頻度順にソートしたものを表 2で示す．形容詞（自立），名詞（形容動詞語幹）にも印象語ではない語が出現頻度上位に含まれていることが確認できる．

そこで，印象語の品詞だけでなく，その前後に共起する品詞に着目し，これらの品詞の出現傾向の偏りによって印象語の判別をする方法を考える．まず，三つの語の並びを考え，それぞれの語に対応する品詞の並びを品詞共起パタンと定義する．このパタンの傾向について以下で分析する．

表 2　印象語である語と印象語ではない語

形容詞（自立）			名詞（形容動詞語幹）		
判定	単語	頻度	判定	単語	頻度
0	多い	68	1	綺麗	57
1	美しい	63	1	きれい	35
1	良い	52	0	大変	23
0	ない	49	0	残念	19
1	涼しい	43	1	有名	19

4.3 印象語の前後の語の品詞

品詞共起パタンを考えるにあたり，まず印象語の前後の語にどのような品詞が出現しているのか，その構成割合を算出した．結果を図 1から図 4に示す．なお，各グラフの内円が品詞の構成比を表し，外円はその品詞を細分化したものの割合を示している．

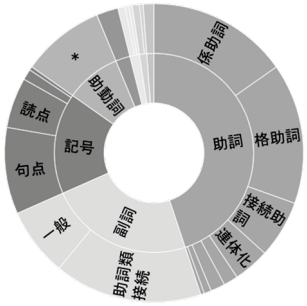

図 1　名詞（形容動詞語幹）が印象語であるとき前に出現する品詞構成割合

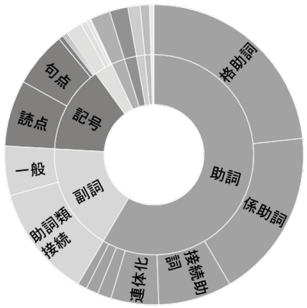

図 2　形容詞（自立）が印象語であるとき前に出現する品詞の構成割合

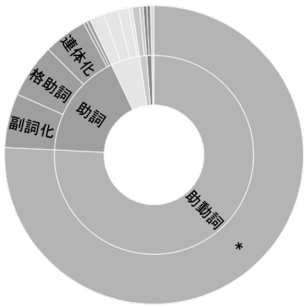

図 3　名詞（形容動詞語幹）が印象語であるとき後に出現する品詞の構成割合

[1] https://www.jalan.net/kankou/

[2] ジャンル分けはじゃらん net のジャンルに基づいている．

[3] https://cran.r-project.org/

[4] http://taku910.github.io/mecab/

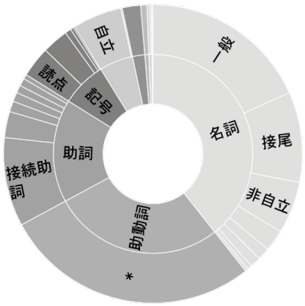

図 4 形容詞(自立)が印象語であるとき
後に出現する品詞の構成割合

これらのグラフから，印象語の前に出現する語の品詞としては，助詞が最も多いことが明らかになった．また，名詞の印象語の後に続く語の品詞は助動詞，形容詞の印象語に続く語の品詞は名詞が多いこともわかった．よって，印象語の品詞によって，品詞共起パタンの出現傾向は異なることになる．印象語を含む高頻度の品詞共起パタンを考えるために，印象語以外の品詞の候補には先に示した構成割合の上位二つの品詞を使用する．品詞共起パタンの構成は，印象語の出現位置によって以下の3種類のケースがある．

① 品詞A+**印象語**+品詞B
② 品詞C+品詞A+**印象語**
③ **印象語**+品詞B+品詞D

たとえば，①のタイプでは，図 1の最も多い品詞は助詞で図 3の最も多い品詞は助動詞となり，品詞Aを助詞，品詞Bを助動詞とすると品詞共起パタンの候補として「助詞+**名詞(形容動詞語幹)**+助動詞」が構成される．表 3に印象語を含む可能性の高い品詞共起パタン候補を示す．このパタンに一つでも当てはまれば印象語とする．

表 3 印象語を含む可能性の高い
品詞共起パタン候補 [5]

種別	品詞共起パタン	種別	品詞共起パタン
①	助詞+**名詞(形容動詞語幹)**+助動詞	②	名詞+助詞+**形容詞(自立)**
①	助詞+**名詞(形容動詞語幹)**+助詞	②	動詞+助詞+**形容詞(自立)**
①	副詞+**名詞(形容動詞語幹)**+助動詞	②	助詞+副詞+**形容詞(自立)**
①	副詞+**名詞(形容動詞語幹)**+助詞	②	記号+副詞+**形容詞(自立)**
①	助詞+**形容詞(自立)**+名詞	③	**名詞(形容動詞語幹)**+助動詞＋名詞
①	助詞+**形容詞(自立)**+助動詞	③	**名詞(形容動詞語幹)**+助詞＋動詞
①	副詞+**形容詞(自立)**+名詞	③	**名詞(形容動詞語幹)**+助動詞＋記号
①	副詞+**形容詞(自立)**+助動詞	③	**名詞(形容動詞語幹)**+助詞＋名詞
②	名詞+助詞+名**詞(形容動詞語幹)**	③	**形容詞(自立)**+名詞+助詞
②	助詞+助詞+**名詞(形容動詞語幹)**	③	**形容詞(自立)**+助動詞+助動詞
②	助詞+副詞+**名詞(形容動詞語幹)**	③	**形容詞(自立)**+名詞+助動詞
②	記号+副詞+**名詞(形容動詞語幹)**	③	**形容詞(自立)**+助動詞+記号

4.4 印象語を含む品詞共起パタンの出現傾向

先に述べた品詞共起パタン候補の出現傾向を明らかにするために，それぞれの出現率を分析し，その上位10位までの品詞共起パタンを表 4に示す．なお，表中の出現率は以下の式で算出した．

$$\text{出現率} = \frac{\text{印象語を含んだ品詞共起パタンの出現頻度}}{\text{品詞共起パタンの出現頻度}} * 100$$

表 4 出現率上位 10 位までの品詞共起パタン

種別	品詞共起パタン	品詞共起パタンの出現頻度	印象語を含んだ品詞共起パタンの出現頻度	出現率
②	記号+副詞+**形容詞(自立)**	41	38	93%
②	動詞+助詞+**形容詞(自立)**	44	39	89%
②	記号+副詞+**名詞(形容動詞語幹)**	24	21	88%
③	**形容詞(自立)**+助動詞+記号	66	56	85%
①	副詞+**形容詞(自立)**+助動詞	58	48	83%
③	**形容詞(自立)**+名詞+助詞	186	153	82%
①	副詞+**形容詞(自立)**+名詞	27	22	81%
①	副詞+**名詞(形容動詞語幹)**+助動詞	64	51	80%
②	助詞+副詞+**形容詞(自立)**	74	58	78%
①	副詞+**名詞(形容動詞語幹)**+助詞	8	6	75%

結果から，出現率だけで見ると「副詞+**名詞(形容動詞語幹)**+助詞」の品詞共起パタンのように，品詞共起パタンの出現頻度が非常に少ないものも上位になってしまう．このような品詞共起パタンはデータ数が少ないため出現率が適切な値を示しているとはいえない．そこで，印象語を含んだ品詞共起パタンの出現頻度と出現率に閾値を設定する．閾値は印象語を含んだ品詞共起パタンの出現頻度が50以上，出現率が70%以上とする．閾値を設定した結果の品詞共起パタンを表 5に示す．

表 5 閾値に基づいた品詞共起パタン

種別	品詞共起パタン	品詞共起パタンの出現頻度	印象語を含んだ品詞共起パタンの出現頻度	出現率
③	**形容詞(自立)**+名詞+助詞	186	153	82%
①	助詞+**形容詞(自立)**+名詞	188	141	75%
③	**形容詞(自立)**+助動詞+助動詞	99	74	75%
②	助詞+副詞+**形容詞(自立)**	74	58	78%
③	**形容詞(自立)**+助動詞+記号	66	56	85%
①	副詞+**名詞(形容動詞語幹)**+助動詞	64	51	80%

5. 印象語抽出方法の検証

ここでは，学習用データに基づき選びだした品詞共起パタンが，検証用データに対して印象語をどの程度的確に抽出できているか検証する．

5.1 品詞共起パタンに基づいた検証

表 5の品詞共起パタンを用いた印象語の抽出結果を表 6に示す．なお，適合率は品詞共起パタンで抽出した印象語候補のうち，実際の印象語が含まれている割合を示す指標であり，再現率は実際の印象語のうち，品詞共起パタンで印象語をどの程度網羅的に抽出できたかを示す指標である[8].

この結果，観光スポット3ヶ所とも適合率，再現率ともにあまり高いとはいえない結果となった．これは品詞共起パタンを多く選びすぎたことにより，ノイズとなる不要な単語を多く抽出してしまったことが原因として考えられる．そこで，選び出す品詞共起パタンを減らすために閾値を見直して再実験する．

表 6 品詞共起パタンに基づく印象語抽出の精度

観光スポット	品詞共起パタンの出現頻度	抽出した印象語の数	実際の印象語の総数	適合率	再現率
上野動物園	77	43	71	56%	61%
ハウステンボス	91	50	90	55%	56%
函館山	70	47	81	67%	58%

[5] 品詞共起パタン内の太文字の品詞は印象語を意味する．

5.2 閾値を再設定した検証

品詞共起パタンを絞り込むため，閾値として出現頻度が50以上，出現率が80%以上と再設定する．このとき選び出された品詞共起パタンを表 7に示す．この品詞共起パタンを用いて，検証用データに対して印象語を抽出した結果を表 8に示す．

表 7 閾値を変えた結果の品詞共起パタン
（出現頻度 50 以上，出現率 80%以上）

種別	品詞共起パタン	品詞共起パタンの出現頻度	印象語を含んだ品詞共起パタンの出現頻度	出現率
③	**形容詞(自立)**+名詞+助詞	186	153	82%
③	**形容詞(自立)**+助動詞+記号	66	56	85%
①	副詞+**名詞(形容動詞語幹)**+助動詞	64	51	80%

表 8 閾値を再設定した場合の品詞共起パタンに基づく印象語抽出の精度

観光スポット	品詞共起パタンの出現頻度	抽出した印象語の数	実際の印象語の総数	適合率	再現率
上野動物園	61	37	71	61%	52%
ハウステンボス	66	40	90	61%	44%
函館山	59	40	81	68%	49%

この結果，適合率は上がったが，再現率が下がってしまった．適合率と再現率はトレードオフの関係であるため，これは当然の結果ともいえる．そこで，2種類の閾値設定のうち，どちらが印象語を抽出するのに適しているかを評価するために適合率と再現率の調和平均であるF値を用いる[8]．F値を算出した結果を表 9に示す．

表 9からすべての観光スポットにおいて，表 5の品詞共起パタンのF値が上回っており，4.4の閾値設定を用いたほうが印象語を適切に抽出することできると結論付けられる．

表 9 印象語抽出精度の F 値

観光スポット	表5の品詞共起パタンのF値	表7の品詞共起パタンのF値
上野動物園	0.58	0.56
ハウステンボス	0.55	0.51
函館山	0.62	0.57

5.3 抽出した印象語の出現頻度

F値を求めたことにより印象語を抽出する適切な品詞共起パタンはわかった．この抽出された印象語が，観光スポットごとの印象を的確に表している言葉であるかを確認するために，抽出された印象語の出現頻度の上位10件までの結果を表 10に示す．

表 10 各観光スポットの印象語上位 10 位の出現頻度

上野動物園			ハウステンボス			函館山		
単語	品詞	頻度	単語	品詞	頻度	単語	品詞	頻度
可愛い	形・自	92	良い	形・自	64	寒い	形・自	57
かわいい	形・自	48	楽しい	形・自	59	良い	形・自	51
楽しい	形・自	45	広い	形・自	48	綺麗	名・形	40
広い	形・自	40	綺麗	名・形	37	素晴らしい	形・自	36
良い	形・自	32	いい	形・自	23	美しい	形・自	33
暑い	形・自	25	寒い	形・自	23	悪い	形・自	24
すごい	形・自	22	素晴らしい	形・自	22	きれい	名・形	23
いい	形・自	16	高い	形・自	20	いい	名・形	21
小さい	形・自	15	美しい	形・自	17	暖かい	名・形	11
珍しい	形・自	13	きれい	名・形	17	高い	名・形	11

※形・自は形容詞(自立)，名・形は名詞(形容動詞語幹)を意味する．

この結果を観光スポット別に考察する．上野動物園では「可愛い」の頻度が最も高いことから，動物園の印象を表していることがわかる．ハウステンボスは他の観光スポットより「良い」，「楽しい」の頻度が高い．ハウステンボスには様々なイベントやアトラクションがあることから，これらの単語がハウステンボスを的確に表していると考えられる．函館山は「綺麗」，「素晴らしい」，「美しい」が続いて出現頻度が高い．これは函館山の夜景が有名なことから，これらの単語が上位に出現していると考えられる．

よって，品詞の細分化と印象語の出現率を考慮した品詞共起パタンによる抽出手法は，各観光スポットの的確な印象を抽出することができるといえる．

6. おわりに

本研究では，形容詞(自立)，名詞(形容動詞語幹)を印象語とし，印象語の出現率を考慮した品詞共起パタンに基づいた抽出を行った．さらに，抽出した印象語の出現頻度を算出することで，抽出した印象語が観光スポットの印象を的確に表しているかを明らかにすることができた．これにより，観光スポットと印象語を対応付けることができ，印象語による情報検索や推薦システムの実現が期待できる．

今後の課題は適合率の向上が挙げられる．今回は「綺麗」や「きれい」などの言葉のゆらぎについて対応していないため，シソーラスを用いて単語の出現頻度の統合処理をし，より正確な出現傾向を分析できるようにするべきである．

参考文献

[1] 株式会社 JTB 総合研究所，スマートフォンの利用と旅行消費に関する調査，2018，https://www.tourism.jp/tourism-database/survey/2018/10/smartphone-2018/，(参照 2019-2-6).

[2] 伊達賢志，北須賀輝明，糸川剛，有次正義，“旅先での観光地選び支援のためのブログを用いた観光地の印象抽出手法”，マルチメディア，分散協調とモバイルシンポジウム 2011 論文集，pp.1566-1579，2011.

[3] 杉原太郎，森本一成，黒川隆夫，“m-RIK:個人の感性特性に対応可能な音楽検索システム”，情報処理学会論文誌 46(7)，pp.1560-1570，2005.

[4] 有本裕亮，芋野美紗子，土屋誠司，渡部広一，“印象語フィルタを用いた映画推薦方式”，情報科学技術フォーラム講演論文集 14(2)，pp.77-78，2015.

[5] 山田忍，“形容詞を用いた観光評判情報の抽出に関する研究”，高知工科大学大学院フロンティア工学コース，2007.

[6] 石田基弘，R によるテキストマイニング入門　第 2 版，森北出版，2017.

[7] 小林雄一郎，R によるやさしいテキストマイニング，オーム社，2017.

[8] 北研二，津田和彦，獅々堀正幹，“情報検索アルゴリズム”，共立出版，2002.