モバイル検索システムのための絵文字に対する意味解析

萩原 正人　水野 貴明
バイドゥ株式会社
{hagiwara, takaaki_mizuno}@baidu.com

1　はじめに

日本のモバイル Web は独自の発展を遂げており，例えば，独自のネットワーク，HTML 仕様，リンク構造を有している．特に，日本独自の文化である「絵文字」はモバイル Web において，様々な用途で幅広く利用されている．例えば，「☀☀今日の天気☀☀」のように，タイトルに対する装飾として用いられる場合や，「帰りに🍺を一杯」のように，語の代替として意味的に用いられる場合がある．

従来のモバイル検索システムでは，このような絵文字の情報を活用しておらず，検索結果にも表示されない．しかし，特に後者の「意味的な」絵文字に関しては，適切にインデックスし検索結果として提示することにより，ユーザー体験を向上させることができると考えられる．具体的には，「帰りに🍺を一杯」という記述を「ビール」や「お酒」というキーワードで検索できれば，より適合した文書を検索できる可能性が高まる．このような絵文字に対する検索を，本稿では絵文字の「意味検索」と呼ぶ．

関連手法として，山本ら(2009)は，絵文字をクエリの一部として入力できる検索手法を提案し，例えば「明るい気分になれるような，かわいらしい映画を調べたい」といった漠然とした情報要求に対して有効であることを示した．しかし，モバイル Web においては，ページ内にも絵文字が頻繁に用いられており(2 節)，この手法では，このような絵文字の情報を活用することができない．また，山本らの手法には，絵文字に対してどのような語が対応するかを記述した辞書を人手で準備する必要があるという問題点がある．絵文字は通信事業者によってそれぞれ多数定義されているうえ(安岡 2007)，モバイル Web においては，絵文字が定義に沿った使い方をされないことも多く(3 節)，このような辞書を作成するコストは高い．語義曖昧性解消(WSD)のタスクにおいても，実際の使用に即さない語義定義が問題となるという指摘がなされている(Véronis 2004)．

そこで本稿ではまず，2 節において，絵文字の使用状況に関する調査結果を紹介し，モバイル Web において絵文字が頻繁に用いられている事実とともに，絵文字が具体的な語の代替として「意味的」に使用されていることを示す．続いて 3 章において，絵文字の意味検索のための準備として，意味的な絵文字の意味や使用例を，コロケーション抽出や分布類似度の手法を用いてマイニングできることを示す．このマイニング結果を用いて，山本らが人手により作成したような絵文字と意味の対応辞書を半自動的に作成することができる．

絵文字の意味検索を実現する際には，主に文脈などの情報を用いて，ある絵文字の出現が装飾的か意味的かどうか，意味的ならば，どの語義に対応するかということを，WSD を用いて判別する．教師有り WSD を適用する場合には，語義情報付きの大規模訓練コーパスが必要になるが，絵文字に関してこのようなコーパスは入手できない．そこで本稿では，半教師あり学習であるブートストラッピングにより，意味的な絵文字の意味を自動推定し，訓練コーパスの作成を支援できることを示す．

2　絵文字使用に関する調査

本節では，絵文字がモバイル Web においてどのぐらい，どのように使用されているかの調査結果について述べる．

2.1　絵文字使用に関する統計

表 1 に，バイドゥモバイル検索[1]docomo 向けインデックスからランダムサンプリングしたモバイル Web ページを対象として，絵文字の使用状況に関する統計を取った結果を示す．ここから，約半分のモバイル Web ページにおいて絵文字が使用されており，かつ，1 ページに含まれている個数も 5~6 個あたりが標準的であることが分かる．

表 1 絵文字使用に関する統計

項目	数値
調査対象モバイルページ数[2]	8,812

[1] http://m.baidu.jp/

[2]「モバイルページ」とは，モバイル端末向けの Web ページのことを指し，バイドゥ独自の分類アルゴリズムによる結果（精度 96%以上）を用いている．今回は docomo 用のモバイルページのみを対象とした．

絵文字を含むモバイルページの割合	46.6%
1 絵文字ページ[3]の絵文字数（平均）	7.6
1 絵文字ページの絵文字数（中央値）	5
絵文字出現率	1 個/246 bytes[4]

次に，絵文字の頻度ランキングを表 2 に示す．数字の絵文字は，accesskey 属性によりテンキーによりアクセスできるハイパーリンクの左側に置かれる用法が大半であり，0はトップに戻る役割として使われることが多い．その他にも (ホームに戻る) や (前に戻る) など，ナビゲーションを補助する目的で使われるものが比較的多いことが分かる．また，(検索)，(投稿日時)，(書き込み)等，特定の機能・情報を表す絵文字も頻繁に用いられる．

表 2 高頻度絵文字上位 20 個

	絵文字		絵文字
1	0 (0)	11	(本)
2	(メモ)	12	(黒ハート)
3	9 (9)	13	(ひらめき)
4	(調べる)	14	(鉛筆)
5	(ハート)	15	4 (4)
6	(ぴかぴか)	16	5 (5)
7	1 (1)	17	3 (3)
8	(時計)	18	(次項有)
9	(家)	19	8 (8)
10	(メール)	20	6 (6)

2.2 使用タイプ

これまで述べたように，絵文字には，装飾的，機能的，意味的な用法があり，本稿では便宜上，この 3 タイプに分類する (表 2)．典型的なモバイル Web ページにおける各タイプの使用例を図 1 に示した．

表 3 絵文字使用の 3 つのタイプ

タイプ	使用法
装飾的	単に装飾する目的で使われる
機能的	直接意味を表さないが，近隣の要素に対して情報を与える
意味的	直接意味を表し，単語の代替として使われる

3 絵文字の意味マイニング

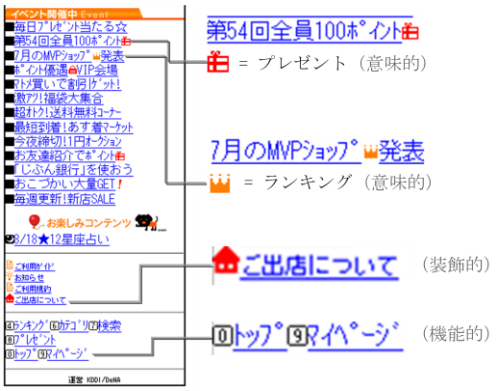

図 1 モバイルページにおける絵文字使用例

前節で述べた 3 つの使用タイプのうち，特に意味的な用法については，その絵文字の実際の意味と，入力されたクエリとをマッチングさせることにより，絵文字の意味検索が実現できると考えられる．

これを実現するために，まず，各絵文字が意味的用法としてどのように用いられているのかを調べる必要がある．絵文字は，本来の定義とは異なった使われ方をすることが多いため，モバイル Web ページから実際の使用例をマイニングする必要がある．例えば，「」という絵文字は，公式には「台風」と定義されているが，実際には「疲れたぁ～」などの文脈で，ネガティブな気分を表す顔文字と同様に用いられる．また，「M」という絵文字も，本来は「地下鉄」の意味であるが，「今日もMで昼」のように，「マクドナルド」の意味で用いられることの方が多い．以下では，コロケーション抽出と，分布類似度の手法を用いて，絵文字の意味をマイニングした結果を述べる．

3.1 コロケーション抽出

モバイル Web ページにおける絵文字の使用法を観察すると，「打ち上げでビール」などのように，左側の文脈（多くは直前）の語に対する装飾的用法として用いられることが多いことが分かる．ここから，コロケーション抽出の手法を用い，絵文字を含む語の連接を抽出することにより，絵文字の意味候補がマイニングできると考えられる．

コロケーション抽出実験には，2 節と同様，バイドゥモバイル検索 docomo 向けインデックスからサンプリングしたモバイル Web ページ 13,847 ページを用いた．また，絵文字の文脈を補充するため，別にサンプリングした約 14 万ページ中の絵文字使用例 228,310 例を追加した[5]．コロケーションの強さの指標としては，

[3] 「絵文字ページ」とは，絵文字を少なくとも 1 個以上含むモバイルページのことを指す．

[4] 文字エンコーディングは UTF8 であり，HTML タグも含む．

[5] ページは，MeCab 0.98pre3 + NAIST jdict 0.6.1 を用いて形態素分割した．英数字の連続は 1 形態素とした．また，HTML タグは，開き・閉じおよびタグの種類の情報のみを用いて，他の語と同様に扱った．

下式によって表される自己相互情報量(Pointwise Mutual Information; PMI)を用いた：

$$\mathrm{pmi}(w_i, w_j) = \log \frac{|w_i, w_j|}{|w_i, *||*, w_j|}$$

ここで，$|w_i, w_j|$は，語w_iと語w_jの共起頻度，$*$はワイルドカードを表す[6].

表3に，絵文字「」と「」に対してコロケーション（形態素前後1~2グラム）を抽出した結果を，PMIの値の大きい順に示した[7]．この結果からも分かるように，両者とも本来の定義とは異なる使われ方をしている．また，絵文字「」「」に対して同様の手法を適用すると，「生ビール」「シャンパン」や「オススメ」「晩ご飯」などのように，絵文字の実際の用法が獲得された．

表4 コロケーション抽出結果

	コロケーション		コロケーション
1	撃沈	1	ド
2	バイト	2	昨日わ
3	考え中	3	わざわざ
4	やばいね	4	そのあと
5	苦手です	5	マック
6	だるかった	6	買って
7	わぁー	7	行った
8	てぇ～	8	行っ
9	ぅぅ	9	行った

3.2 分布類似度

コロケーション抽出の手法を用いることにより，絵文字の意味・使用法を部分的に抽出できることが分かったが，「帰りに」を一杯」などの，文脈中に語義が出現しない意味的用法に対しては依然として語義を獲得できない．

そこで，「文脈の類似した語は，意味も類似している」という分布類似度(Lin 1998)を用いることにより，対象の絵文字と類似した意味を持つ別の絵文字や語を自動獲得できると考えられる．実験には，3.1節と同様のデータおよび前処理を用いた．語に対する文脈として，語の前後2形態素を用い，PMIにより重みづけた．語と語の類似度には，下式で定義されるJaccard係数を用いた：

$$\mathrm{sim}(w_i, w_j) = \frac{\sum_c \min\left(\mathrm{pmi}(w_i, c), \mathrm{pmi}(w_j, c)\right)}{\sum_c \max\left(\mathrm{pmi}(w_i, c), \mathrm{pmi}(w_j, c)\right)}$$

ここで，$\mathrm{pmi}(w_i, c)$は，語w_iに対する文脈cのpmiの重みである．この手法を用い，絵文字「」と「」に対して類似絵文字・語を出力した結果を，類似度の大きい順に表4に示した．

表4 コロケーション抽出結果

	の類似絵文字・語		の類似絵文字・語
1	(目がハート)	1	「top」
2	(ぴかぴか)	2	「トップ」
3	(うれしい顔)	3	(1)
4	(パー)	4	(9)
5	(黒ハート)	5	(ビル)
6	(上向き矢印)	6	(0)
7	(ウッシッシ)	7	「店舗」
8	(ほっとした顔)	8	(位置情報)
9	(るんるん)	9	(次項有)

結果から，「」に対しては，主に文末に置かれることによりポジティブな感情を表す絵文字が，「」に対しては「トップ」や「」「」など，実際の意味的・機能的な用法に即した類似絵文字・語が獲得できている事がわかる．また，分布類似度を用いることにより，用法の傾向によって絵文字や語のクラスタリングができ，装飾的・意味的な絵文字の判別に有効であると考えられる．

4 絵文字のインデックス化

絵文字をインデックスするにあたって，Webページ内における絵文字の表現方法はキャリアによって異なっているため，それぞれのページのエンコーディングおよび対象とするキャリアを知る必要がある．しかし，収集されたモバイルページの多くは，ページコンテンツ以外には，どのキャリア向けの絵文字を含んでいるかを判別できる情報を持たない．そこで，エンコーディングを判定し，絵文字のコード情報を利用してキャリアを判別し，両者を統一してインデックスした．

各キャリアはウェブ上で利用できる絵文字表現を複数有しており，キャリアが公式に情報を公開していないものも含め，3キャリア19タイプの絵文字表現を認識している．例えば，docomoには，を表す表現として4種類がある(表6).

表6 docomoのの絵文字表現

タイプ	実際のコード
ShiftJISバイナリ表現	F89F
UTF-8バイナリ表現	EE 98 BE
数値実体参照(16進数/UTF-8)	
数値実体参照(10進数/Shift_JIS)	

例えば，というタグ列は，<a> という2単語の列として扱われる．

[6] なお，PMIが負の値となった場合，0とした．

[7] HTMLタグを含むもの，記号のみを含むものは除外した．

ただし，19 タイプの絵文字コードの一部は領域が重複しているため，表現や絵文字の利用頻度を用いて推定を行った．統一化に際しては emoji4unicode[8]をベースとし，独自の内部表現を利用している．

5 ブートストラップ法による意味判別

3 節の結果により，絵文字（タイプ）に対して意味をマイニングできたが，意味検索の際には，各出現（トークン）に対して具体的な語義を割り当てる必要がある．このタスクは，語の出現に対して語義を割り当てる WSD によって定式化できる．

WSD にはいくつかのアプローチがあるが，特に教師有り WSD を適用する場合には，語義情報付きの大規模訓練コーパスが必要になる．しかしながら，絵文字に関してこのようなコーパスは入手できない．

そこで本研究では，ブートストラップ法 (Komachi et al. 2008)のアプローチを用い，少数の教師データから，大量の用例に対する語義判別を試みた．具体的には，*Espresso* アルゴリズム (Pantel and Pennacchiotti 2002) を用い，少数のシードから，同一の用例の他の出現を繰り返し的に獲得する．ここで，シードとしては，絵文字のある意味的用例，例えば，「ビール」の意味で用いられている「」の出現（トークン）を与える．*Espresso* アルゴリズムの概略を以下に示す：

1. シード（用例）を与える
2. シードと共起する信頼度の高い文脈パターンを抽出した後，汎用パターンを削除する[9].
3. パターンと共起する信頼度の高いインスタンスを抽出した後，汎用インスタンスを削除し，シードに追加する．
4. 条件を満たすまで，2〜3 を繰り返す．

実験では，3 節で補充用に用いた約 14 万ページ中の絵文字使用例 228,310 例と，その前後 1 行ずつ抽出したものをコーパスとして用い，上記 2~3 のステップを 5 回繰り返した．繰り返し毎に，シードを 5 個ずつ追加し，最終的に得られたインスタンス中に残った絵文字の使用例について，その用法の精度・再現率を評価した．その結果を表 5 に示す．

表 5 絵文字意味自動判別の結果

対象	用例	シード	精度・再現率
＝ビール	7 例／345 個	3 例	71.4%／71.4%
＝プレゼント	28 例／1270 個	5 例	100%／53.6%

[8] http://code.google.com/p/emoji4unicode/

[9] 「汎用パターン」「汎用インスタンス」は，それぞれパターン，インスタンス中での相対出現頻度が 0.6%以上であるものとした．

この結果から，少数のシードから，他の用例が獲得できていることが分かり，抽出された用例を人手により確認することにより，大量の用例を半自動獲得できる可能性があることが示唆される．

6 おわりに

本稿では，絵文字の意味検索のために，まず，絵文字の使用状況を調査し，絵文字がモバイル Web において予想以上に幅広く使われていることを示した．また，コロケーション抽出および分布類似度の手法を用いて，絵文字の意味をマイニングできることを示した．また，ブートストラップ法により，少数の正解データから大量の用例を獲得できる 可能性を示した．本研究の成果は，バイドゥモバイル検索に既に部分的に応用されており，例えば「六本木ヒルズ」というクエリで，意味的な用法のみを検索することが可能である．

今後の課題として，Komachi et al. (2008) は，グラフカーネルを用いると意味ドリフトを抑えながら高精度に WSD が適用できることを示したが，このような高度な WSD の手法も比較・検討する．また，Yarowsky (1995)は，One sense per discourse，すなわち，語は，単一の文書中では一つだけ語義を持つ傾向があるという仮説を提唱したが，このような，近接文脈以外の情報なども有効であるか検討してみたい．

参考文献

山本千尋，安田宜仁，別所克人，内山俊郎，内山匡．クエリとして絵文字を受け付ける情報検索．人工知能学会第 23 回全国大会論文誌，2B2-2， 2009.

安岡孝一．ケータイの絵文字と文字コード．情報管理，科学技術振興機構，2007 年 5 月号，p.71．2007.

Dekang Lin. Automatic retrieval and clustering of similar words. *Proc. of COLING/ACL*, pages 786-774, 1998.

Mamoru Komachi, Taku Kudo, Masashi Shimbo and Yuji Matsumoto. Graph-based Analysis of Semantic Drift in Espresso-like Bootstrapping Algorithms. *Proc. of EMNLP*, pp.1011-1020, 2008.

Patrick Pantel and Marco Pennacchiotti. Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations. *Proc. of ACL*. pp. 113-120, 2002.

Jean Véronis. Hyperlex: lexical cartography for information retrieval. *Computer Speech and Language*, Vol. 18, No. 3, pp. 223-252, 2004.

David Yarowsky. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. *Proc. of ACL*, pp. 189–196, 1995.