絵文字を考慮したテキスト解析の研究

山本千尋† 別所克人† 内山俊郎† 内山匡†
日本電信電話株式会社 NTTサイバーソリューション研究所†

1. はじめに

携帯メールによるコミュニケーションの特徴の一つとして、絵文字の利用が挙げられる。絵文字とは、携帯キャリアが独自に用意している特殊文字である。最近では、絵文字は携帯メールにとどまらず、携帯対応のブログや、twitterなどでも利用可能となっている。

絵文字は、文中で何らかの意味を表現する場合があるため、文章を構成する要素としてテキスト解析結果に反映される必要がある。しかし出現箇所によって文中における役割や語義を複数持つ曖昧性があるため通常の語と同様なテキスト解析を行うことができない。そこで、絵文字を含む文のテキスト解析を可能にするためには、絵文字の曖昧性を解消する必要がある。

曖昧性の解消では、単語の語義を辞書などによってあらかじめ定義し、曖昧性を持つ語の語義を、定義された語義から決定するため、単語の語義をあらかじめ定義した辞書が必要である。

絵文字は、“絵”であるという特徴から、特に曖昧性を持つ性質があるため、日々語義が変化していると考えられる。語義の変化に対応しない場合、辞書の再現率に影響を与えると考えられるが、人手で変化に対応し追加を行うのは非常にコストがかかる。

そこで、本研究では、絵文字の曖昧性の解消を目的とし、絵文字の語義を自動で抽出する手法を提案する。

2. 関連研究

既存の語義の自動抽出に関する研究では、語の分布を利用した方法や、グラフに基づく方法がある。語の分布を利用する方法では、文脈内に出現する単語から構成されるベクトルをクラスタリングすることで単語の語義を抽出する研究がある [1]。グラフに基づく方法では、共起語をノードとしてグラフで表現し、クラスタリングを行うことで単語の語義を抽出する研究がある[2]。

3. 絵文字の語義の抽出

絵文字の語義を、自動で抽出する手法を提案する。

本研究では、絵文字の「お日様☀がでてきたね。」のように文中の内容語（名詞、動詞、形容詞）が示す意味を、絵文字を用いてもう一度文に添加する性質（以下、内容添加）に着目し、絵文字と共起する内容語から、語義の抽出を行う。内容添加では、絵文字の語義である内容語が文中に出現するため、絵文字と、絵文字と共起する内容語との意味的な距離が非常に近くなると考えられる。

そこで、絵文字と共起する内容語の中から、類似する内容語をまとめた語クラスタを生成し、絵文字との意味的な距離を用いて信頼度の高い語クラスタを抽出する手法をとる。

(1). 共起語の抽出

上述の内容添加の性質により、絵文字の語義が、絵文字が出現する文中で共起する内容語に含まれる場合がある。そこで絵文字の異なりごとに、絵文字が出現する一文中の内容語を、絵文字の語義候補として抽出する。

(2). 語クラスタの生成

(1)で抽出された、絵文字の異なりごとの語義候補には、例えば、「☀」では、{晴れ, 天気}など天気に関する語、{夏, 夏休み, 日焼け}など夏に関する語など、上位概念に共通点があるものが含まれている。そこで、類似している語を集約し絵文字の異なりごとに、上位概念に共通点がある語クラスタを生成する。

類似する語を集約するために、語義候補に意味を表すベクトルを与え、ベクトル間の距離によってクラスタリングを行う。語義候補にベクトルを与える方法として、概念ベース技術を用いる[3]。概念ベース技術とは、単語と単語の意味属性とのコーパス中における共起頻度に基づき単語の概念ベクトルを生成するものである。与えられた概念ベクトルは、クラスタリング手法であるウォード法によってクラスタリングを行う[4]。

(3). 信頼度の高い語クラスタの抽出

(2)で生成された語クラスタには、絵文字が表現する意味と関係のないものも多くある。絵文字が表現する意味と関係のない語クラスタを排除するために、概念ベース技術によって生成された絵文字の概念ベクトルと各語クラスタとの

A method for processing a text with pictograms.
Chihiro YAMAMOTO, Katsuji BESSHO, Toshio UCHIYAMA, Tadasu UCHIYAMA
NTT Cyber Solutions Laboratories, NTT Corporation

距離と、語クラスタに含まれる共起語の共起回数を用い、絵文字が表現している意味に関係がある語クラスタを抽出する。

概念ベースによって生成された絵文字の概念ベクトルは、絵文字と共起する単語の意味属性によって構成されているため、絵文字の語義を複数含んだ形で生成される。絵文字の概念ベクトルと、各語クラスタの重心との距離が閾値以上離れている語クラスタを排除することで、絵文字ベクトルに含まれる、絵文字を構成する様々な語義のいずれとも離れた語クラスタを排除することができる。

語クラスタの中には、絵文字とほとんど共起しない語のみによってなるクラスタも存在する。内容添加の性質より、共起回数が多い語が含まれるクラスタの方が、絵文字の語義を表現するクラスタとして信頼度が高いといえるため、絵文字との共起回数の平均が閾値未満の語クラスタは排除する。

4. 実験

絵文字入りブログ34万エントリに200回以上出現する絵文字138個を対象に、提案手法を用いて、語義の抽出を行った。その結果、85個の対象絵文字について語義が抽出された。抽出語義数は、2～88個だった。

また、抽出された語義について、絵文字を作成した携帯キャリアが絵文字に与えた文字表現である“タイトル”と比較した。

4-1. 抽出語義の適合率の評価

絵文字85個それぞれに対して抽出された語義が絵文字を表現する語義であるかについて、主観評価により人手で確認した。主観評価の基準としては、絵文字から容易に連想できるものを正解とした。その結果、抽出した語が8割以上適合している絵文字が53個(63%)、5割以上8割未満が適合している絵文字が7個(8%)、5割未満が適合している絵文字が25個(29%)であった。

4-2. 考察

表1に本提案手法によって抽出された語義の例を示す。

表1　抽出された語義

絵文字	タイトル	抽出された語義
🍺	ビール	ビール、晩酌、発泡酒、忘年会、飲み会、歓迎会、新年会など
☀	晴れ	晴れ、快晴、朝焼け、起床、早起き、猛暑、真夏日、夕日、夕焼け　など

例に示すように、提案手法では、絵文字「🍺」について、文字表現として与えられたタイトルである「ビール」以外にも、「忘年会」、「飲み会」などが抽出された。また、絵文字「☀」についても、タイトルである「晴れ」以外にも「早起き」など朝に関連するものや、「真夏日」など夏に関連するものが抽出された。

この結果より、本手法では、絵文字の言語表現として与えられたタイトルの語義の他に、ユーザが実際に絵文字に与えている語義を抽出できることが分かった。

語義が抽出されなかった絵文字と、適合率が低かった絵文字について分析したところ、1.「⑵番目」や「いやだ‼」のように絵文字が通常の文字と同じように用いられるもの(20%)、2.「パーティー♪」や「遅刻💦」のように、絵文字が感情を表現するもの(52%)、3.その他(28%)であった。絵文字の感情を表す語義では、今回語義が抽出できた絵文字についても抽出が難しかった。感情を表す語義を抽出する方法としては、絵文字と共起する感性語（形容詞）を語義として抽出する方法が考えられる。例えば、絵文字「♪」と共起する感性語を抽出すると｛楽しい、嬉しい、美味しい｝となり、これを用いることで、文中で表している感情表現を抽出できると考えられる。

5. まとめ

本稿では、絵文字の曖昧性の解消を目的とし、絵文字の曖昧性解消のために必要となる絵文字の語義を自動で抽出する手法を提案した。今後は、今回作成した辞書を用いて、絵文字の曖昧性の解消手法について検討を行う。

参考文献

[1]Hinrich Shutze, “Automatic word sense discrimination”, Computational Linguistics, Vol.24, No.1, pp.97-123, 1998

[2] Beate Dorow, Dominic Widdows, and Katarina Ling, “Using curvature and markov clustering in graphs for lexical acquisition”

[3]別所克人,内山俊郎,内山匡,片岡良治,奥雅博:“単語・意味属性間共起に基づくコーパス概念ベースの生成方式”, 情報処理学会論文誌,Vol. 49 No.12 pp.3997-4006, 2008

[4] 神嶌敏弘,“データマイニング分野のクラスタリング手法(1)－クラスタリングを使ってみよう！－,”人工知能学会誌, Vol.18, No.1, pp.59-65, 2003