4T-05

# 日本語と英語における絵文字の使われ方の差異の調査

内藤 大輝 †　　矢吹 太朗
千葉工業大学 社会システム科学部 プロジェクトマネジメント学科 ‡

## 1　序論

日本から生まれた絵文字が今では Unicode にも含まれており，世界中で使われている．

同じ絵文字でも文章によって様々な意味を持つことがある [1]．そのような違いの一因として，絵文字が使われる文章の言語の違いがあると思われる．そこで本研究では，異なる言語における，絵文字の意味の違いを調査する．

本研究では，word2vec[2, 3] のベクトルを，単語や絵文字の意味と見なす．日本語の文章と英語の文章を収集し，word2vec によって単語や絵文字をベクトルで表現する．そうしてできるベクトルを分析することで，個々の絵文字が持つ意味の，言語（日本語と英語）における違いを発見する．

## 2　目的

Deepmoji[4] で用いられていた，よく使われる 64 種類の絵文字を含むデータを Twitter から取得し，ここの絵文字の，日本語の文章中での意味と英語の文章中での意味を比較する．

## 3　手法

次の手順で調査する．

1. Twitter から絵文字を含むツイートを日本語と英語の 2 種類を取得する．
2. 日本語のデータを分かち書きをする．
3. 絵文字のツイートを用いて word2vec のモデルを作成する．
4. word2vec のモデルを用いて主成分分析する．
5. word2vec のモデルを用いてクラスター分析する．
6. 絵文字のコサイン類似度を出力し，日本語のモデルと英語のモデルで比較する．

## 4　結果

絵文字を含むツイートから word2vec のモデルを日本語と英語の 2 種類作成した．

日本語と英語のモデルを用いた主成分分析の結果を図 1 と図 2 に示す．日本語と英語のモデルを用いたクラスター分析の結果を図 3 と図 4 に示す．

表 1 に絵文字のコサイン類似度を出力して日本語文中と英文中で比較した際に意味が異なった絵文字を示す．例えば，U+1F648 では日本語では照れを表し，英語では愚かなという意味を表している．

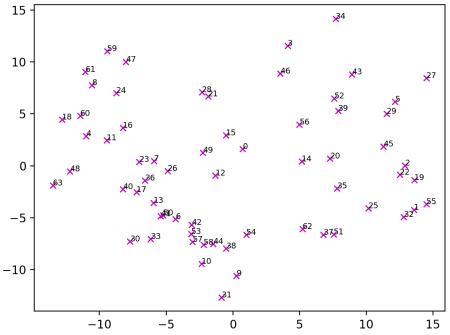

図 1　日本語のモデルにおける主成分分析の結果

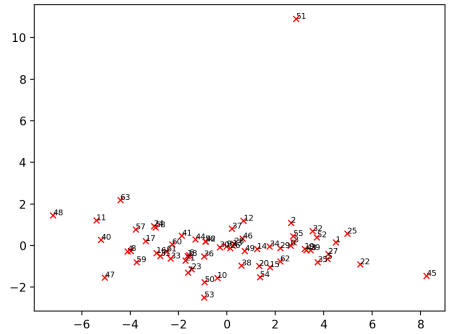

図 2　英語のモデルにおける主成分分析の結果

Investigation of the influence of failure of web service based on tweet analysis.

† Daiki NAITO (s1642091ez@s.chibakoudai.jp)
‡ Department of Project Management, Faculty of Social Systems Science, Chiba Institute of Technology.

図 3　日本語のモデルにおけるクラスター分析の結果

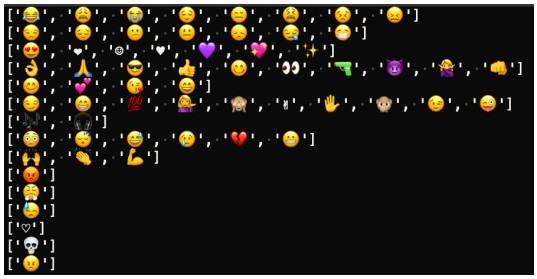

図 4　英語のモデルにおけるクラスター分析の結果

表 1　日本語文中と英文中で意味が異なる絵文字

| コードポイント | 絵文字 | 日本語 | 英語 |
|---|---|---|---|
| U+1F648 | 🙈 | 照れ | silly |
| U+1F480 | 💀 | キャラクター | spooky |
| U+1F648 | 🙈 | 含み笑い | think |
| U+1F64F | 🙏 | お願い | bless |
| U+1F590 | 🖐 | はい | Raise |
| U+1F4AA | 💪 | 頑張る | sports |
| U+1F608 | 😈 | キャラクター（複数） | |
| U+1F4AF | 💯 | 満点 | （複数） |

## 5　考察

日本語のモデルを主成分分析した図 1 を見ると全体的に分散していることがわかる．また，英語のモデルを主成分分析した図 2 を見ると下に要素が密集していることがわかる．

絵文字のコサイン類似度を出力し確認したところ，分散している単語は意味が異なっており，密集している単語同士は似通ったものになっていた．

また，日本語のモデルをクラスター分析した図 3 は感情ごとに分かれていることが見て取れるが英語のモデルをクラスター分析した図 4 は 5 行目や 9 行目以降で意味のまとまりがわからないものとなっている．

このことから実際によく使われており，意味が確立している単語とあまり使われていない単語があるのではないかと推測される．

日本語では絵文字のコサイン類似度を出力すると他の絵文字が出てくることが多くあった．日本語では絵文字を用いて文章の内容を表しているが，英語では単語を言い換えているため絵文字が連続して使用されないのではないかと考えられる．

## 6　結論

Twitter からデータを取得し word2vec でモデルを 2 種類作成した．それらを用いて，日本語と英語における絵文字の使われ方の差異を調査した．その結果，日本語での絵文字の使われ方と英語での絵文字の使われ方に違いがあることを確認した．今後，厳密に違いを調査することにより翻訳やコミュニケーションの助けになることが期待できる．

## 参考文献

[1] Heeryon Cho, 稲葉利江子, 石田亨, 高崎俊之, 森由美子. 絵文字コミュニケーションにおけるセマンティクス. 情報処理学会研究報告. ICS, [知能と複雑系], Vol. 145, pp. 1–8, oct 2006.

[2] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, Vol. abs/1301.3781, , 2013.

[3] 斎藤康毅. ゼロから作る Deep Learning 2—自然言語処理編. オライリー・ジャパン, 2018.

[4] Bjarke Felbo, Alan Mislove, AndersSøgaard, Iyad Rahwan, Sune Lehmann. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017.