

絵文字解釈語を用いた意味的関連度の提案

Semantic Relevance Measure Using Pictogram Interpretation Words

Cho Heeryon¹ 石田亨¹ 稲葉利江子² 高崎俊之³ 森由美子³ 神田智子⁴
Heeryon Cho Toru Ishida Rieko Inaba Toshiyuki Takasaki Yumiko Mori Tomoko Koda

京都大学 大学院情報学研究科 社会情報学専攻¹

Department of Social Informatics, Graduate School of Informatics, Kyoto University

独立行政法人情報通信研究機構 知識創成コミュニケーション研究センター 言語グリッドプロジェクト²

Language Grid Project, National Institute of Information and Communications Technology

特定非営利活動法人パンゲア 京都 R&D センター³

Kyoto R&D Center, NPO Pangaea

大阪工業大学情報科学部情報メディア学科⁴

Department of Media Science, Faculty of Information Science and Technology, Osaka Institute of Technology

1 はじめに

世界の子供たちが絵文字メッセージをやり取りして親睦を深める絵文字コミュニケーションシステムがある[1]。システムで使われる絵文字は創作絵文字であり、感情や文化など非形式的な情報を伝え、多様な解釈を持ちうる。

絵文字メッセージを作成する際に、メッセージ作成者は、登録絵文字の中から必要な絵文字を探す。一つの絵文字が多様な解釈を持ったり、複数の絵文字が同じ解釈を共有したりするので、どの絵文字を使えばよいか迷う場合がある。このとき、ある解釈を持つにふさわしい絵文字を、支持される解釈の順に並べることができれば、メッセージ作成者が参考にすることができる。

我々は、ある解釈を検索語として投げ、絵文字に付与されている解釈語を手がかりに、その検索語にふさわしい絵文字を順位付けするための尺度を提案する。

2 意味的関連度

任意の検索語が与えられたとき、個々の絵文字がその検索語にどれくらいふさわしいかを意味的関連度 (semantic relevance, $SR(w_i, e_k)$) として定義する。このとき絵文字は、解釈語のリストと、解釈語ごとにその解釈語を付与した人間の数を頻度として持つ。

ある絵文字における、ある解釈語の出現率 (その解釈語の頻度を、その絵文字のすべての解釈語の頻度合計で割った値) は、その絵文字がどれくらい多くの人たちに、その解釈語で支持されているかを表す。

絵文字 e_k の解釈語を w_1, w_2, \dots, w_n とし、その解釈語の出現率を $P(w_1|e_k), \dots, P(w_n|e_k)$ とする。このとき、 w_i を検索語としたときに、絵文字 e_k を提示することが妥当であることを示す最も簡単な指標は $P(w_i|e_k)$ になる。しかし、このままでは解釈語間の類似度は考慮されない。これを解決するために、解釈語間の類似度を定義する。これを用いて、 w_i を検索語として用いたときに、絵文字 e_k を提示することが妥当であることを示す指標 $SR(w_i, e_k)$ を次で計算できる。

E_i は、 w_i が解釈語となる絵文字の集合を、 $|E_i \cap E_j|$ は、 w_i と w_j 両方を解釈語として持つ絵文字の要素数を、 $|E_i \cup E_j|$ は、 w_i か w_j のどちらか一つでも解釈語として持つ絵文字の要素数を表す。

表1 検索語「すやすや」の意味的関連度 (SR)

絵文字	SR	解釈語と頻度
	0.645	おやすみ [17] ねる [9] すやすや [7] おやすみなさい [6] 寝る [4] ねむる [3] 睡眠 [3] ぐーぐー [2] ねてる [2] 夢 [2] 夢の中 [2] 夜 [2] ...
	0.193	おやすみ [44] おやすみなさい [16] 寝言 [5] 夢 [5] ねごと [4] 寝る [3] 睡眠 [2] ねむる [1] ゆめ [1] 家の夜 [1] 寝ている [1] ...
	0.141	夜 [63] よる [16] おやすみ [3] 深夜 [2] おやすみなさい [1] にわとり [1] コケコッコ [1] 就寝 [1] 朝 [1] 夜中 [1] 夜明け [1] 夕方 [1] ...

$$SR(w_i, e_k) = \sum_j P(w_j|e_k) |E_i \cap E_j| / |E_i \cup E_j| \quad (1)$$

意味的関連度 (SR) を計算する絵文字検索システムのプロトタイプを実装し、絵文字 120 個の解釈語リストと頻度情報を与え、提案尺度が有効であるかを実験した。表 1 に、検索語「すやすや」を投げたときの検索結果を示す。二番目のコラムに SR 値を記す。検索結果の一番上の絵文字は「すやすや」という解釈語を持つが、それ以外の絵文字は持たない。しかしどれも眠っている様子を描いており、「すやすや」という解釈を持つにふさわしい絵文字といえる。

3 おわりに

絵文字コミュニケーションシステムに使われる創作絵文字は、多様な解釈を持ったり、同じ解釈を共有したりするので、どの絵文字を使えばよいか迷う場合がある。意味的関連度を提案することで、ある解釈を持つにふさわしい絵文字を順位付けできた。本稿に掲載された絵文字の著作権は特定非営利活動法人パンゲアに帰属します。

参考文献

- [1] Toshiyuki Takasaki. PictNet: Semantic Infrastructure for Pictogram Communication. In *Proc. of the 3rd Int'l WordNet Conf. (GWC-06)*, pp.279-284, Jan. 2006.