形態素情報を用いた 系列ラベリングによる顔文字抽出

高嶋　浩平† 　森　康久仁† 　松葉　育雄†
† 千葉大学大学院融合科学研究科

1. はじめに

近年, SNS の普及によりこれらに投稿される記事を解析して活用する動きが活発化している. しかし, これらの記事は現代的な表現などで崩れた日本語表記であることが多く, 従来の形態素解析器では正しく処理できないことがある[1]. 崩れた日本語表記の一つである顔文字に関しては, 顔文字辞書によって形態素として処理する以外には特別な処理がなされていないことがほとんどで, しばしば誤った処理結果が出力されてしまうことがある.

先行研究[2]ではこの問題を解決するため, 系列ラベリングの手法で記事から顔文字を自動で抽出し形態素解析器の顔文字辞書の充実を図っていたが, 抽出に失敗してしまう顔文字もあった. 失敗したのは主に括弧の外に手などの文字があるものや, 括弧のない顔文字などである. 以下に例を示す.

o(*⌒―⌒*)o　　∩∧ω∧∩

そこで本研究では, 過去に抽出できなかった顔文字を抽出するための改善策を提案する.

2. 顔文字抽出の手法

2.1. CRF によるラベル付け

Conditional Random Field(CRF)は系列ラベリングの手法の一つである[3]. 入力された文の各文字に対し前後の文字の情報からラベル付けをしていく. 使用するラベルは B(顔文字の 1 文字目), I(顔文字の 2 文字目以降), T(それ以外の文字), EOS(文の終端)の四種類で, B と I のラベルの連続を顔文字とする.

2.2. 素性

素性とは, ラベル付けをする際に利用する特徴のことである. 先行研究[2]では文字そのものに加えて文字の種類も利用している. 各文字を C(日本語, アルファベット), N(数字), S(それ以外の文字), EOS(文の終端)の四種類に分類する. 表 1 に素性と正解ラベルの例を示す.

本研究では, 素性を文の品詞と形態素の情報に変更し, より文章に近い形で顔文字抽出をする手法を提案する. 表 2 は「美味しいね^_^」という文を形態素解析し, その時に正解ラベルを付与した例である. 形態素解析器には juman を利用した[4].

表 1 先行研究の素性と正解ラベル

位置	文字	文字の種類	正解ラベル
1	で	C	T
2	す	C	T
3	(	S	B
4	・	S	I
5	ω	S	I
…	…	…	…

表 2 提案手法の素性と正解ラベル

位置	形態素	品詞	正解ラベル
1	美味しい	形容詞	T
2	ね	終助詞	T
3	^	未定義語	B
4	_	未定義語	I
5	^	未定義語	I
6	EOS	EOS	EOS

3. 実験と結果

TwitterAPI によって収集したツイートを用いて, 実験を行った. ラベルを付与した 2000 件のツイートを学習データとして CRF によって学習させ, 500 件のツイートを評価データとして顔文字を抽出した. 実験は先行研究の手法と提案手法の両方で行い, 結果を比較した.

顔文字抽出の F 値を測ったところ, 先行研究の手法では 87.4%なのに対し, 提案手法では 87.5%と先行研究とほぼ同じだが, 先行研究では抽出に失敗した, 括弧の外側に文字のある顔文字や括弧のない顔文字の抽出に成功することがあった. 以下に例を示す.

m(_ _)m　　(^^)v　　Σ(゜д゜lll)　　^o^

4. まとめ

今回の手法により先行研究で抽出できなかった顔文字も抽出できるようになる可能性を示した. 今後は素性などの改良を行い, 抽出精度を高めていきたい.

参考文献

[1] 利根川翔, 筧捷彦, "崩れた表記に対応する日本語形態素解析器の開発, "情報処理学会 75 回全国大会, 分冊 2, no.1Q-4, pp.115-116, 2013.

[2] 渡邉謙一, 高橋寛幸, 但馬康宏, 菊井玄一郎, "系列ラベリングによる顔文字の自動抽出と顔文字辞書の構築, "言語処理学会 第 19 回年次大会 発表論文集, 分冊 1, no.P6-13, pp.866-869, Mar. 2013.

[3] 高村大地, "系列ラベリング, "自然言語処理シリーズ 1 言語処理のための機械学習入門, 5章, コロナ社, 東京, 2010.

[4] 黒橋・河原研究室, "JUMAN-KUROHASHI-KAWAHARA LAB , " 京 都 大 学 , http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN, 参照 Nov. 12, 2015.