文字種と画数を用いた新若者語の抽出

秋田 恭佑　　松本 和幸　　北 研二
徳島大学 工学部 知能情報工学科
{akita-kyosuke}@iss.tokushima-u.ac.jp

1　はじめに

近年，スマートフォンの爆発的な普及に伴って，Twitter や facebook を始めとする SNS やブログの利用者数が増加している傾向にあり，中でも，スマートフォンで SNS やブログを利用している人のうちの大半が１０代～３０代の若者である．そのため，SNS 上での情報交換の場では若者特有の言葉 (若者語) が用いられる事が多い．若者語とは，若者が日常的に用いている言葉のことであり，非常に流行に乗りやすく日々新たに生成されては消えているため，網羅するのが困難であるといった特徴がある．

また，若者語は，一部のコミュニティ内でしか通用しないものもある．コミュニティ外の人がブログを読んだりする場合に若者語の意味を理解することが難しく，どの語が若者語であるかもわからないといった状況も考えられる．

本研究では若者語は日常的に使用されやすい語を用いて生成されるという仮定に基づいて，語の表層的な特徴より文中の若者語を抽出することを検討する．文中からの抽出には条件付き確率場（以下 CRF）を用いて確率モデルで文字列が若者語である確率の高い言葉を推定することにより行う．

2　関連研究

本研究と同様に若者語を対象としている研究には，周辺情報から若者語を抽出する研究 [1] や若者言葉を含んだ発話文から感情推定を行う研究 [2] があるが，どちらも Web サイトに登録されている若者語を対象としているため，最新の若者語に対応できていない．また，前者の研究ではカタカナのみで構成されている若者語しか考慮していないため，網羅性に欠ける．本研究ではカタカナ以外の文字種で構成されている若者語にも対応した抽出方法を検討する．

3　若者語の特徴分析

本稿では，若者語の抽出実験をするにあたって，Web から若者語を含む文を収集し，その文中に含まれている若者語を人出で抽出している若者語コーパス [3] と，語感の辞典 [4] に掲載されている単語に Casual か Formal のタグを付与した語感データベースを使用する．前者のコーパスには 20500 の若者語を含む文と約 26000 個の単語 (同一語を含む) が収録されている．表 1 に若者語コーパスの収録文の例を記述する．

表 1: 若者語コーパスの収録文の例

ジベタリアンって単語に笑っちゃいました ww
善意から来てるのがわかるだけ絶妙に ウザい（笑
ありござ で?す (o) ノ
まいしてう ＼ (o) ／同盟 まいしてう
ルビ子さんのレロレロ顔も ウケる～

また，このコーパスに収録されている単語の文字種パターン別の頻度を測ってみると，もっとも多かったのは，カタカナで 8077 回 (30.91 %) であった．次に多かったのは，カタカナ＋ひらがなであり，続いて，ひらがなのみ，カタカナ＋漢字，漢字のみであった．カタカナによる若者語が頻出することから，文字種は若者語らしさを表す情報となり得ると考えられる．

また，事前に語感データベースを使用し，若者語コーパス内の若者語の画数とフォーマルの度合いを分析した．その結果，若者語は画数の少ない Casual タグが付与された文字より比較的画数の多い Formal タグの付与された文字を使用していることが明らかになったため，漢字の若者語の抽出に画数は有効であると考えられる．

本実験では，この事前情報と抽出結果からどの文字種パターンが精度良く抽出できるかについても実験する．なお，以降，カタカナを「カナ」，ひらがなを「かな」，漢字を「漢」，英字を「英」，数字を「数」，

記号を「記」と表記する．

4 CRFでの推定方法

本稿では，若者語の抽出にCRFを用いるが，CRFを計算し，結果を推定するツールとしてCRF++[1]を採用した．学習データは文を文字単位に分割し，その１文字ずつに対して若者語かそうでないかのラベルを付与したものである．また，ラベルの他にも文字種と画数の情報を付与する．同様にテストデータも用意し，テストデータの正解ラベルとCRFにより推定されたラベルから再現率，適合率，F値を算出し，考察する．以下に再現率，適合率，F値の計算式を示す．なお，下式でRは推定した文字列から正解した数を，Cはテストデータ内の若者語の数，Nは推定した文字列の総数を表している．

$$Recall = \frac{R}{C} \quad (1)$$

$$Precision = \frac{R}{N} \quad (2)$$

$$F-measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (3)$$

また，本研究では新しい若者語の抽出および若者語の箇所の推定が目的であるため，テストデータと学習データに同じ若者語は含めない．今回の実験では，学習・テストに用いる情報として，１）文字のみ，２）文字+文字種，３）文字＋文字種＋画数の３通りについて再現率，適合率，F値を算出している．

また，本実験では，学習データを１）500文，２）1000文，３）5000文で実験しており学習データ数が結果にどう影響するのかについても結果を記す．今回，使用するテストデータ内の文字種パターン別単語頻度の多い順に上位5パターンを表2に示す．

表2: テストデータ内若者語の文字種パターン別頻度

文字種パターン	頻度 (%)
カナ	99(35.61)
かな	48(17.27)
カナ+かな	39(14.03)
カナ+漢	33(11.87)
漢	23(8.27)

次に，CRF++で使用する学習データの例を図1に示す．左から，文字，文字種，画数，ラベルを表している．

貰	漢字	12	O
っ	ひらがな	1	O
た	ひらがな	4	O
女	漢字	3	B
子	漢字	3	I
力	漢字	2	E
高	漢字	10	O
そ	ひらがな	3	O
う	ひらがな	2	O

図1: 学習データの例

また，CRF++で使用した素性テンプレートを図2，3，4に示す．

```
U00:%x[-2,0]
U01:%x[-1,0]
U02:%x[0,0]
U03:%x[1,0]
U04:%x[2,0]
U05:%x[-2,0]/%x[-1,0]/%x[0,0]
U06:%x[-1,0]/%x[0,0]/%x[1,0]
U07:%x[0,0]/%x[1,0]/%x[2,0]
```

図2: 文字のみを素性とした場合

```
U00:%x[-2,0]
U01:%x[-1,0]
U02:%x[0,0]
U03:%x[1,0]
U04:%x[2,0]
U05:%x[-2,0]/%x[-1,0]/%x[0,0]
U06:%x[-1,0]/%x[0,0]/%x[1,0]
U07:%x[0,0]/%x[1,0]/%x[2,0]
U08:%x[-2,1]/%x[-1,1]/%x[0,1]
U09:%x[-1,1]/%x[0,1]/%x[1,1]
U10:%x[0,1]/%x[1,1]/%x[2,1]
```

図3: 文字と文字種を素性とした場合

[1] http://crfpp.googlecode.com/svn/trunk/doc/index.html

```
U00:%x[-2,0]
U01:%x[-1,0]
U02:%x[0,0]
U03:%x[1,0]
U04:%x[2,0]
U05:%x[-2,0]/%x[-1,0]/%x[0,0]
U06:%x[-1,0]/%x[0,0]/%x[1,0]
U07:%x[0,0]/%x[1,0]/%x[2,0]
U08:%x[-2,1]/%x[-1,1]/%x[0,1]
U09:%x[-1,1]/%x[0,1]/%x[1,1]
U10:%x[0,1]/%x[1,1]/%x[2,1]
U11:%x[-2,0]/%x[-2,1]/%x[-2,2]
U12:%x[-1,0]/%x[-1,1]/%x[-1,2]
U13:%x[0,0]/%x[0,1]/%x[0,2]
U14:%x[1,0]/%x[1,1]/%x[1,2]
U15:%x[2,0]/%x[2,1]/%x[2,2]
```

図 4: 文字と文字種と画数を素性とした場合

5 結果

本実験での結果を以下に記す．CRF で推定した若者語が，正解と完全一致した場合の再現率，適合率，F 値は表 3，4，5 のようになった．なお，以降，若者語の情報として文字のみを用いた場合を C，文字＋文字種の場合を C+K，文字＋文字種＋画数を用いた場合を C+K+S と表記する．

表 3: 学習データ数別の再現率

学習データ数	C	C+K	C+K+S
500	19 %	32 %	36 %
1000	24 %	28 %	33 %
5000	26 %	28 %	29 %

表 4: 学習データ数別の適合率

学習データ数	C	C+K	C+K+S
500	64 %	55 %	54 %
1000	66 %	63 %	64 %
5000	76 %	81 %	79 %

表 5: 学習データ数別の F 値

学習データ数	C	C+K	C+K+S
500	30 %	40 %	43 %
1000	35 %	39 %	44 %
5000	38 %	42 %	42 %

また，文字種パターン別の抽出頻度についても，学習データ数を 500 文に設定した際に，C 型と C+K 型ではカナ＋かなの若者語の抽出数が 21 個から 29 個に増えた．また，C+K 型と C+K+S 型では，変化はあまり見られなかったが，C+K+S 型は唯一，漢字の若者語を抽出することができた．学習データを 1000 文にすると，C+K 型と C+K+S 型ではカナの若者語の抽出数が 26 個から 38 個に増えた．しかし，学習データを 5000 文にすると，抽出数が全体的に低くなり，カナの若者語の抽出数は C+K+S 型より C+K 型のほうが多かった．

6 考察

抽出に成功した若者語の例を表 6，誤抽出の例を表 7 に示す．

表 6: 提案手法による正解例

アブラも、ヤバイ
きしょい!きしょすぐる
自己中 で自意識過剰で
しないためにも ハデ婚 をすれば
ニコ生 は現実世界で相手に
ドタキャン されてしまいました

表 7: 提案手法による誤抽出例

下旬に インパー クしようと (インパーク)
なかなか エグい死 に方を (エグい)
トゥルトゥルの アブラ も
それに 壁紙もゲット 出来るよう (ゲット)
チョコ をパケ買いさせる
モリパパ!ハデ婚バンザイ!!(ハデ婚)
優秀 プチプラチーク です (プチプラ)

文字に加えて文字種や画数などを用いた場合と用いなかった場合を比較すると，文字以外の情報を加えた

場合で再現率の改善がみられたが，画数の有無で抽出精度に変化はあまりみられなかった．また，文字種パターン別の抽出頻度をみると，文字の情報として画数を用いることにより漢字で構成されている若者語の抽出精度が増加している．

また，学習データを 5000 文にした場合，画数を文字の情報として用いた場合と用いない場合では後者の方がカナで構成された若者語の抽出精度が良い結果が得られた．これは，カタカナに画数のばらつきが少ないためであると考えられる．

また，正解と同様に間違いも多数見られた（表 7）．中には,「ハデ婚」など一方では抽出に成功した若者語が異なる文中では誤抽出されていた語も存在した．また，誤抽出された文字列の中には,「インパー」や「プチプラチーク」など元の若者語が特定できる，あるいは若者語を含んでいるようなものもあった．今後これらの部分一致した文字列を対象に，分析を進めていくことで，抽出の手がかりを見つける必要がある．

7 おわりに

本研究では， 文字種と画数を素性とした CRF による未知若者語の抽出を提案した. 学習データ数が 500 文と少ない場合では， 画数を素性として用いることが有効であることが示された. また，今回の実験では，完全に抽出できた若者語を対象としていたため，今後は， 部分的に抽出できた文字列に後処理を加えるなどして， 抽出精度を高めたい.

謝辞

本研究は，科学研究補助金（若手研究（B），23700252）により行われた．

参考文献

[1] 松尾 朋子, 安藤 一秋 , テンプレートを用いた Web からの若者言葉の抽出手法の検討 , 第 11 回情報科学技術フォーラム , 2012

[2] 松本 和幸, 任 福継 ， 感情推定における若者言葉の影響 , 言語処理学会第 17 回年次大会 , pp.846-849 (2011)

[3] Kazuyuki Matsumoto, Kenji Kita and Fuji Ren , Emotion Estimation from Sentence Using Relation between Japanese Slangs and Emotion Expressions , 26th Pacific Asia Conference on Language, Information and Computation , pp.377-384 (2012)

[4] 中村 明 , 語感の辞典 , 岩波書店 , 2010