D-008

若者言葉の自動抽出に用いたテンプレート改良に関する検討
Consideration of Improving Templates for Automatic Extraction of Young People's Words

松尾 朋子†　　安藤 一秋‡
Tomoko Matsuo　　Kazuaki Ando

1．はじめに

近年，一般家庭にもインターネットが普及し，誰でも簡単に Web 上で情報発信が可能となった．特に若者（10～20 代）は，日常会話で利用している若者特有の言葉（若者言葉：10～20 代までの若者が Web 上でよく使用する言葉）をブログや Twitter などでも使う傾向がある．若者言葉は，使用期間が短く，新しい言葉が造られやすい特徴がある．また，若者言葉は日本語の文法や規則から逸脱したものもあり，若者言葉に親しみのない世代には，若者言葉の意味を理解できない場合がある[1-3]．例えば，若者言葉に親しみのない世代が若者言葉で書かれたブログを読む場合，若者言葉の意味を理解できず，ブログの内容を理解できない場合がある．そこで，ブログの内容を理解するためには，若者言葉の意味が記載されている書籍や Web サイト，検索エンジンを利用して若者言葉の意味を調べる必要がある．しかし，若者言葉の意味が記載されている書籍や Web サイトでは若者言葉の収集や意味の調査を人手で行っているため網羅性に欠ける．

そこで本研究では，Web から若者言葉を自動収集し，それらの意味を推定する手法の実現を目的とする．著者らは，先行研究[1]で，既存の若者言葉（種言葉）の直前・直後の形態素に注目したテンプレートを用いて，種言葉と同種のカタカナ若者言葉が自動抽出できる可能性について検討した．その結果，種言葉の直前・直後の形態素のみを利用したテンプレートでは，種言葉と同種の若者言葉を抽出することは難しいことが分かった．そこで，本稿では，若者言葉が含まれるブログ記事の書式調査とテンプレートの改良法について検討する．

2．先行研究の概要

著者らの先行研究[1]の概要について説明する．

2.1　テンプレートの構築

既存のカタカナ若者言葉を種言葉とし，種言葉の直前・直後に出現する形態素を基にテンプレートを構築する方法を説明する．まず，種言葉に対し，「直前の形態素」+「種言葉」+「直後の形態素」というパターンを利用して，それぞれの種言葉に対する直前・直後の形態素を収集する．次に，「種言葉」の部分をワイルドカードに置き換えたパターン「直前の形態素」+「*」+「直後の形態素」をテンプレートとして利用し，* の位置に出現するカタカナ言葉を若者言葉候補として収集する．

2.2　テンプレートを利用した抽出手法

テンプレートを用いて若者言葉候補を抽出する手法の手順を以下に示す．

① Web 検索を用いたテンプレートの選別
② テンプレートを用いたワイルドカード検索
③ 若者言葉候補の収集
④Web 辞書を用いたフィルタリング

2.3　年代別検索を用いた若者言葉の判定

2.2 の手法で得られた若者言葉候補（c_words）に対し，goo の年代別検索（10 代～50 代）を用いて，若者言葉としての妥当性を判定する手法を以下に示す．

①「の」を用いた各年代における登録記事数の推定
② c_words を用いた年代別検索によるヒット数の取得
③「の」と c_words のヒット数による使用割合の計算
④ 若者（10 代～20 代）と他年代（30 代～50 代）の c_words の使用割合による判定

2.4　評価結果の概要

167 語の種言葉から作成した 3,716 個のテンプレートを用いて抽出手法の評価を行った．抽出対象には，テンプレートをクエリーとして Yahoo!ウェブ検索 API で取得したスニペット群を利用した．実験の結果，434 語が若者言葉として抽出され，人手による判定で 40.6%（176/434）が若者言葉と判定された．残りは，人名やキャラクター名などの固有名詞，スペルミス，抽出ミスなどであった．

次に，評価実験に用いたテンプレートの有用性を評価するために，テンプレート作成に用いた種言葉と同じ種類（オノマトペ，省略，合成語，造語の 4 種）の言葉が抽出される割合を分析した．分析対象は，抽出実験によって得られた 549 語（重複を含む）である．分析の結果，549 語の内，テンプレート作成に用いた種言葉と同じ種類の若者言葉が抽出できた割合は，16.2%（89/549）と低く，種言葉と異なる種類の若者言葉が多数抽出されていることがわかった．次に，テンプレート作成に用いた種言葉の内，新しい若者言葉を抽出できた割合を調査した結果，35.9%（60/167）となった．また，60 語の種言葉の内，種言葉と同じ種類の若者言葉が抽出できた割合は，45.0%（27/60）であった．

以上より，種言葉の直前・直後の形態素のみを利用したテンプレートでは，文脈的な制約が弱いため，種言葉と同種の若者言葉を抽出することは難しいことが分かった．そこで，以降では，文脈情報をテンプレートに取り入れるため，係り受け構造の利用を検討するための調査とテンプレートの検討を行う．

3．ブログ記事の書式調査

係り受け解析を行うためには，ブログ記事を一文単位に分割する必要がある．文分割の際，一般的には句点や記号，改行情報を利用する．しかし，若者言葉を利用する年代は，書式の自由度も高く，また，携帯端末を利用して記事を書くことも想定できるため，句読点の欠如や絵文字の利用，不適切な箇所での改行など書式的な問題が懸念される．そ

†香川大学大学院工学研究科　Graduate school of Engineering, Kagawa University
‡香川大学工学部　Faculty of Engineering, Kagawa University

こで，若者が書いたブログ記事の行末書式の調査と文分割・併合の可能性について調査する.

3.1 行末書式の調査

若者言葉を含む文からテンプレートを生成するため，調査対象は既存の若者言葉を含む記事のすべての行末とし，それらの行末書式（句点，記号，文字）について調査する．調査対象には，Yahoo!ブログの学校カテゴリより 2011 年 11 月に収集した若者言葉を含むブログ記事 631 記事と Yahoo!のブログ検索 Web API より 2012 年 2 月に収集した若者言葉を含むブログ記事 52,604 記事を利用する．

調査の結果，最も多かったのは，顔文字や「!」，「?」などの記号で終わる行末で，71.46%であった．続いて，文字で終わる行末が 18.71%，句点で終わる行末が 9.83%であった．行末に記号が多数出現した理由として，「！」や「？」，顔文字などの記号を句点の代替に利用するユーザが多いことが挙げられる．また，行末が文字で終わる理由として，次の行へ文が続く状況以外にも，句点や記号を使用せず，改行を文末記号に利用するユーザの存在も確認できた．行末が文字で終わる場合は，改行が文末記号になっている傾向が多くみられた．

以上を基に，任意の記事の書式を目視調査した結果，ブログに若者言葉を多用するユーザは，文末記号として記号を多用し，句点をあまり利用しない傾向が確認できた．また，行末が記号で終わる記事が 71.46%と多いことから，文末後に改行する傾向も見られた．

3.2 文分割・併合の可能性調査

ブログ記事には，一行に複数文が書かれる場合や複数行で一文が書かれる場合があるため，一行に複数文の場合は分割し，複数行で一文の場合は併合する必要がある．そこで，文末以外に句点が存在するブログ記事を対象に，一行に複数文が書かれる文を分割できるか，複数行で一文が書かれる文を併合できるのかを調査する．対象は，3.1 のデータから任意抽出した 60 記事である．

調査の結果，一行に複数文が書かれる記事は，60 記事中 49 記事（81.67%），複数行で一文が書かれる記事は，60 記事中 28 記事（46.67%）であった．一行に複数文が書かれる場合，図 1 のように文の終わりが句点や「!」，「?」を使用していることが多い傾向があった．複数行で一文が書かれる文には，図 2 のように次の文に続くことを示す読点が書かれている場合や文末が文字終わりとなっている場合があった．さらに全体の傾向として，行末に文末としての句点が使われるより，行末以外の位置に文末としての句点が使われることが多いことも確認できた．

以上の結果より，文末以外に句点や「!」，「?」が存在した場合，一行に複数文書かれている可能性があるため，句点や「!」，「?」で文の分割が可能であると考えられる．また，複数行で一文となる文の併合は，行末が読点終りの場合はよいが，文字終りの場合は誤って併合する可能性がある．そこで，文末が読点の場合のみを次の行の文と併合することで，誤って合併する可能性が低くなると考えられる．

> 何で見て見ぬフリ？お前のせいで絡まれてんだぞ？マジキチ乙ぅー。↵
> 愚痴でーす(http://blogs.yahoo.co.jp/impreza_0113/23257108.html)

図 1　一行に複数文が書かれている例

> アイスバーンにもなっていなくて、普通に雨が降ったあとみたいになっているけど、↵
> 歩道はすごいカチコチに凍っていて、何回もこけそうになった。↵
> 今日って祝日なのに…
> (http://blogs.yahoo.co.jp/chami7201schloss/21138958.html)

図 2　複数行で一文となる例

4. 文脈情報を加味したテンプレートの検討

種言葉の直前・直後の形態素のみを利用したテンプレートでは，種言葉と同種の若者言葉を抽出することは難しい．そこで，係り受け構造を利用することで文脈情報を加味したテンプレートの構築について検討する．

同じ文脈で利用される若者言葉は，動詞の場合，同じ格要素を取ることが多いと考えらえる．また，名詞の場合，同じ意味の動詞に係り，同じ格要素になりやすく，他の格要素も類似していると考えられる．そこで，まずは動詞に係るタイプのカタカナ若者言葉を種言葉に設定し，若者言葉が係る動詞とその動詞が取りうる格要素の情報を用いてテンプレートを構築することを検討する．対象とする格要素は，鈴木ら[2]と同様，文脈の特徴が表れやすいと考えられるヲ格，デ格，ニ格に注目してテンプレートを構築する．

以下に，テンプレートを構築する手順の骨格を示す．

① 「種言葉+格助詞（ヲ，ニ，デ）」でブログ検索し，ヒット数を取得
② ①で取得したヒット数が最大の「種言葉+格助詞」を用いてブログ記事を収集
③ 種言葉を含む一文を係り受け解析し，「種言葉+格助詞（=格要素）」+「動詞」となる「動詞」の頻度をカウントすると共に，その他の格要素を取得
④ で最多頻度となった「動詞」と格要素などの情報を用いてテンプレートを構築

今後は，上記手法の詳細について検討を継続する．

5．おわりに

本稿では，若者言葉を含むブログ記事の書式調査と種言葉と同じ意味の若者言葉を抽出するために，係り受け構造に注目し，文脈情報を用いたテンプレートについて検討した．

今後は，テンプレートを構築するための具体的な方法を考案する．そして，構築したテンプレートを用いてカタカナ若者言葉を自動抽出し，テンプレートの性能および妥当性を評価する．

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，第 2 分冊，pp.79-80，2012.
[2] 鈴木雄登，笹野遼平，高村大也，奥村学，“リムる・ドヤる・ポジる・パフェる―Web を用いたカタカナ動詞の言い換え・語源の獲得―”，情報処理学会研究報告，2012-NL-209(8)，pp.1-7，2012.
[3] 秋田恭佑，松本和幸，北研二，“文字種と画数を用いた新若者語の抽出”，文言語処理学会第 19 回年次大会(NLP2013)予稿集，2013.