D-001

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

Consideration of a Method for Extracting Young People's Words from the Web using Templates

松尾 朋子† 安藤 一秋‡
Tomoko Matsuo Kazuaki Ando

1．まえがき

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）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，新しい言葉が常に作られる特徴がある．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．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．また，更新頻度も少ないため，新しい若者言葉の意味を調べることもできない．

若者言葉を対象とした研究として，若者言葉を既知語へ復元する研究[1]や若者言葉の感情推定を行う研究[2]がある．しかし，これらの研究では若者言葉が記載されている書籍や Web サイトから収集した若者言葉を対象としているため，最新の若者言葉については考慮していない．

そこで本研究では，Web から若者言葉を自動収集し，意味を推定する手法の実現を目的とする．本稿では，研究の初期段階として，Web から若者言葉を自動抽出する手法について検討する．なお，本稿では，若者言葉の文字種パターンを調査した結果，最も多いことが確認された，カタカナ表記の若者言葉を対象とする．

2．若者言葉の定義

本研究では，10～20 代までの若者が Web 上でよく使用する言葉を若者言葉と定義する．若者言葉は，「夫婦（若者言葉ではカップルという意味）」のように一般に使用される言葉（以下，一般言葉）と表記が同じでも異なる意味で使用される場合がある．本稿では，抽出が容易であると考えられる一般言葉と表記が重複しない若者言葉を抽出対象とする．また，書き言葉としての若者言葉には発音が困難な文字（例：エ゜）を含むものが存在するが，若者言葉は主に若者の日常会話の中で生成されるという性質を考慮し，まずは発音できる若者言葉を対象とする．

以上より，本稿では，一般言葉と表記が異なり，発音が可能な若者言葉を最初の抽出対象とする．

†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

3. 若者言葉の分析

3.1　若者言葉の収集

若者言葉の抽出手法を検討するために，若者言葉の特徴を分析する．本稿では，人手で収集した若者言葉を公開している「若者言葉辞典～あなたはわかりますか？～」，「みんなで国語辞典！」，「日本語俗語辞書」の 3 つのサイトから機械的に収集した 3,290 語の若者言葉を分析対象とする．収集した若者言葉には，一般言葉と同一表記の言葉（例：夫婦（カップルという意味））が含まれる．そこで，一般の辞書を用いて，一般言葉と同一表記の若者言葉を取り除く．使用する辞書は，三省堂 Web Dictionary，goo 辞書，Weblio 辞書，Yahoo!辞書，kotobank の 5 つである．

5 つの辞書のいずれかに登録されている言葉を取り除いた結果，1,108 語の若者言葉が得られた．取り除かれた 2,182 語には，一般言葉と同一表記の語の他に，世間に広く認知されている若者言葉（例：イケメン）が含まれるが，分析に与える影響は少ないと考える．一方，1,108 語の中にも一般に使用される語（例：×２（バツ２））が存在した．そこで，一般に使用される語を手作業で削除した結果，最終的に，1,079 語（例：マジで，リア友，ガン見）が得られた．以下，これらの若者言葉を利用して分析を行う．

3.2　文字種パターンの分析

3.1 で収集した若者言葉 1,079 語を用いて，若者言葉を構成している文字種パターンを調査する．また，各文字種パターンの出現頻度についても調査する．なお，調査対象の文字種は，ひらがな，カタカナ，漢字，アルファベット，記号とする．

調査の結果，文字種パターンは全部で 74 種類存在した．また，若者言葉の文字種パターンで最も多いものは，カタカナのみで構成された若者言葉であり，若者言葉 1,079 語中 167 語（15.48%）であった．次に，漢字のみで 146 語（13.53%），カタカナ+漢字で 105 語（9.73%），カタカナ+ひらがなで 90 語（8.34%），ひらがなのみで 89 語（8.25%）の順であった．そのため，まずはカタカナ表記の若者言葉を自動抽出する手法について検討する．

4．若者言葉候補を抽出するためのテンプレートの定義

カタカナで構成された 1,079 語の若者言葉に対する事前調査により，若者言葉の直前や直後には若者言葉に関連のある言葉が出現することがわかった．そこで，1,079 語の若者言葉に対し，「直前の言葉」+「若者言葉」+「直後の言葉」というパターンを利用して各若者言葉に対する直前・直後の言葉を収集する．その後，「若者言葉」の部分をワイルドカードに置き換えたパターン「直前の言葉」+

「*」+「直後の言葉」をテンプレートとして，* 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．

この手法により，テンプレート作成に利用した若者言葉と同じ文脈で使用される若者言葉候補が抽出できるため，テンプレート作成に利用した若者言葉と意味的に何らかの関係がある若者言葉が収集できると考えられる．

5. カタカナ表記の若者言葉の自動抽出手法

5.1 カタカナ表記の若者言葉候補の抽出

既存の若者言葉と同じ文脈で使用される新しい若者言葉を Web から抽出するために，4. で定義したテンプレートを用いた抽出手法を提案する．以下に，若者言葉候補を抽出するための手順を示す．

① テンプレートの収集

カタカナ表記の若者言葉を含むブログ記事から若者言葉を含む一行を抽出し，テンプレートを収集する．例えば，ブログ記事から「テンションアゲアゲで行った」が抽出できたとする．抽出した一行を形態素解析し，「テンション/アゲアゲ/で/行った」と分割する．「カタカナで構成された若者言葉」は「アゲアゲ」であるため，「直前の言葉」の部分は「テンション」，「直後の単語」は「で」となり，「テンション」+「アゲアゲ」+「で」を得る．その後，「アゲアゲ」をワイルドカードに置き換えることにより，テンプレートは「テンション」+「*」+「で」となる．

② テンプレートの選別

「直前の言葉」と「直後の言葉」が付属語の場合，若者言葉との関係性が弱いため，テンプレートとしては利用できない．そこで，「直前の言葉」と「直後の言葉」のどちらか一方が内容語の場合，テンプレートとして利用する．

得られたテンプレートを用いて，Web で検索し，ヒット数を得る．その後，ヒット数が 1 以上のテンプレートを収集する．Web 検索には，Yahoo!API を用いる．

③ ワイルドカード検索

②で選別したテンプレートを用いて，ワイルドカード検索し，検索結果からスニペットを収集する．

④ カタカナ表記の若者言葉候補の収集

③で得られたスニペットから，ワイルドカード部分にあたる言葉を抽出する．抽出対象とする言葉は，長音記号「ー」を含むカタカナで構成された若者言葉である．

⑤ Web 辞書を用いたフィルタリング

④で得られたカタカナ表記の言葉が一般言葉である場合，フィルタリングし，一般言葉でない場合，若者言葉候補として収集する．フィルタリングには，三省堂 Web Dictionary，Weblio 辞書，Yahoo!辞書，kotobank の 4 つの辞書を用いる．

5.2 年代別検索を用いた若者言葉の判定

5.1 で得られた若者言葉候補が若者言葉であるのかを goo の年代別検索を用いて判定する．以下に年代別検索（10 代～50 代）を用いた若者言葉の判定法の流れを示す．

① 各年代における登録記事数の推定

「の」を用いて年代別検索し，各年代のヒット数を得る．得られたヒット数を各年代の登録記事数とする．

② 若者言葉候補を用いた年代別検索

若者言葉候補を用いて年代別検索し，各年代でのヒット数を得る．

③ 使用割合の計算

①で得られた「の」のヒット数と②で得られた若者言葉候補のヒット数を用いて年代別の割合を求める．

④ 若者言葉の判定

若者（10 代～20 代）と他年代（30 代～50 代）の若者言葉候補の使用割合を比較して判定する．若者の使用割合が大きい場合，若者言葉として出力する．

6. 抽出実験

167 語のカタカナ表記の若者言葉から作成した 3,716 個のテンプレートを用いて，新しい若者言葉を抽出する実験を行い，提案手法の妥当性を確認する．

実験の結果，テンプレートを用いて得られた 1,209 語の若者言葉候補に対し，年代別検索で若者言葉の判定を行った結果，434 語が抽出された．そこで，抽出された若者言葉の妥当性を検証するため，434 語中 217 語を任意で選択し，人手で分析する．

分析の結果，217 語中 92 語（42.40%）が若者言葉であった．残りの 125 語（57.60%）は，人名やキャラクター名，スペルミスなどであった．若者言葉と判定した 92 語には，「メタキン（ゲームキャラクター「メタルキング」の略称）」や「テキトー（適当をカタカナに置き換え，口語的に表現した言葉）」のように若者言葉の可能性がある言葉が 24 語（26.09%）含まれている．次に，若者言葉の可能性がある言葉を除いた 68 語（73.91%）について確認する．68 語中 6 語（ガノタ，ガリマッチョ）は，テンプレートを作成するために利用した若者言葉であった．17 語（イケボ，コミュ）は，ニコニコ大百科やはてなキーワードに意味が登録されていた．したがって，テンプレートを用いて新しく抽出できた若者言葉は 45 語（デキメン，ウザキャラ）であった．

7. おわりに

本稿では，テンプレートを用いて，Web からカタカナ表記の若者言葉を抽出する手法について検討した．実験の結果，若者言葉の抽出精度は 42.40%であった．抽出精度があまり高くないため，今後は，固有名詞やスペルミスを取り除く手法と構文構造を用いたテンプレートなどについて検討する．また，テンプレート利用の有効性を検証するため，テンプレート作成に利用した若者言葉と抽出された若者言葉の意味関係を調査する．そして，最終的には，若者言葉に意味推定手法を考案する．

参考文献

[1] 原田俊信，山本義人，久保村千明，佐々木洋輔，亀田弘之，“若者語処理システムの評価”，電子情報通信学会技術研究報告，AS-4-2，"S-37"-"S-38"，2006

[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