若者言葉の意味と感性を考慮した標準語への変換手法

松本 和幸[1]　土屋誠司[2]　吉田 稔[1]　北 研二[1]
[1] 徳島大学大学院ソシオテクノサイエンス研究部
[2] 同志社大学理工学部インテリジェント情報工学科
{matumoto, mino, kita}@is.tokushima-u.ac.jp
stsuchiy@mail.doshisha.ac.jp

1　はじめに

自然言語処理の分野において新語などの辞書に登録されていない未知語の適切な解析を目的とした様々なアプローチが提案されている [1, 2]．しかし，これらの研究の多くは固有名詞（人名，施設名，製品名，etc.）の処理が主なターゲットであり，若者言葉のように明確に定義することが困難な語については研究例が少なく，今後の進展が期待されている．

テキスト中の俗語の意味や感性を理解できるか否かが，とくに若者が記述した文章からの意見・評判分析の精度に影響を与えると考える．また，若者言葉のなかには，標準語を言い換えた表現も多く存在するため，若者言葉を標準語に変換（復元）することができれば，意見・評判分析や感情認識の精度改善に貢献できる．

本研究では，若者言葉文を標準語文に変換するために，「意味」と「感性」の 2 つの基準から変換候補を選択する手法を提案する．「意味」と「感性」を考慮することにより，意味情報を保持したまま，適切な感性を表現する文に変換できると考える．

たとえば，「明日もバイト入ってて タヒる わ」という若者言葉を用いた文から，標準語を用いた文への言い換えを考える．意味のみを考慮した場合，標準語を用いた文は，「明日もバイト入ってて 死ぬ わ」といった表現となる．「死ぬ」という言葉に置き換えることにより，本来の意味の「死ぬ」といった，より深刻な状況を連想させてしまうため，感性を適切に表現できているとはいえない．このため，「明日もバイト入ってて しんどい わ」といった文に変換できることが望ましい．このように，意味のみではなく感性を考慮すれば，意味的には遠いが感性的には近い変換候補の取りこぼしを減らすことができると考える．

2　若者言葉の印象分析

若者言葉は，仲間内の会話において，過激な内容の発言を柔らかい印象に変化させたり，言葉では表現しにくい状況などを伝える際に臨場感を持たせたりすることなどによく用いられる．一方で，標準語は，不特定多数の人に発話の意図や意味を正しく伝えることを第一目標としている．このため，意味的に同じか，あるいは類似する若者言葉と標準語が必ずしも同一の印象を与えるとは限らない．このことから，それぞれの若者言葉がどのような印象を持っているか，また，それらの印象が標準語とどのような違いがあるかを分析する必要がある．

2.1　若者言葉の印象アンケート

若者言葉が与える印象を，アンケートに対する複数の被験者の回答を分析することで調査する．本調査では，若者言葉感情コーパス [3] に含まれる若者言葉と，ニコニコ大百科[1] において，語釈文に若者言葉，ネットスラング，隠語という表記のある見出し語を合わせて 671 語選定し，ランダムで 2 等分し，被検者 1 名あたり約 300 語について回答する形式とした．各語に対し少なくとも 2 名以上の被験者が回答するようにした．アンケートの回答には専用の回答ツールを用いて，各表現に対し，16 種類の印象対（感情対も含む）を設け，各々50 段階で評価する．アンケート結果を 16 次元のベクトルに変換し，各若者言葉間のユークリッド距離を算出した．この結果について，一部の若者言葉に関して多次元尺度構成法に基づき 2 次元座平面上に配置したものを，図 1 に示す．

図中，同じマーカで示す若者言葉は，印象ベクトルを用いてクラスタリングをおこなった際に同じクラスタ

[1] http://dic.nicovideo.jp/

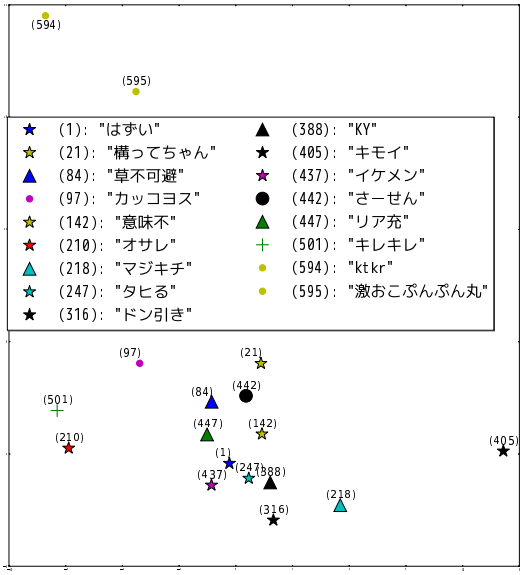

図 1: 2 次元座標平面上における配置

に属したものである．クラスタリングツールには bayon ver.0.1.0 [2] を用いた．クラスタリングアルゴリズムとして Repeated Bisection を用い，分割ポイントの閾値を 1.0 に設定したところ，クラスタ数 k は 41 となった．クラスタリング結果と図における位置関係は必ずしも一致していない．この図から，似通った印象を与える若者言葉も存在する一方で，他の語とは異なる独特な印象を与える若者言葉も存在していることがわかる．

2.2 若者言葉と標準語の印象比較

つぎに，若者言葉に対応する標準語との印象の比較分析をおこなうため，前述の印象評価アンケートにおいて用いた 671 語のなかから，以下の 2 つの条件に当てはまる若者言葉の抽出をおこなった．

- 同一表記の語（意味が異なるものも含む）が既存の標準語辞書には登録されていない
- 意味が同一または類似する表現が標準語辞書に登録されている

本研究では既存の標準語辞書として日本語 WordNet[4]，日本語語彙大系 [5]，分類語彙表 [6]，EDR 概念辞書 [7] の 4 つの辞書を用いた．抽出された語は 154 語となった．本節では，この 154 語の若者言葉に注目してみる．本来ならば標準語についても，若者言葉と同様の印象評価のアンケート分析をおこなう必要がある．しかし，同じ意味の標準語でも異なる表記で記述されることで異なる印象を与えることがあると考えられる．そのため，あらゆる表記を網羅したアンケートを実施することは不可能であることから，本研究では標準語に関しては positive/negative/netural の印象についてのみを対象に比較分析をおこなう．

若者言葉と変換対象の標準語の positive/negative/neutral の組み合わせを集計した結果を図 2 に示す．この結果から若者言葉と標準語との印象が一致する割合が 66.8%であることがわかった．一致しない場合もある程度みられることから，若者言葉から標準語へ変換することで印象が変化してしまう（positive/negative が反転する）可能性がある．

印象が一致しない組み合わせにおいて，若者言葉が positive，標準語が negative の場合がもっとも多く，18 組あった．つまり，標準語では negative な意味にとらえられがちの語でも，若者言葉で表すことにより，positive な印象を与えることができるケースが多くあるといえる．

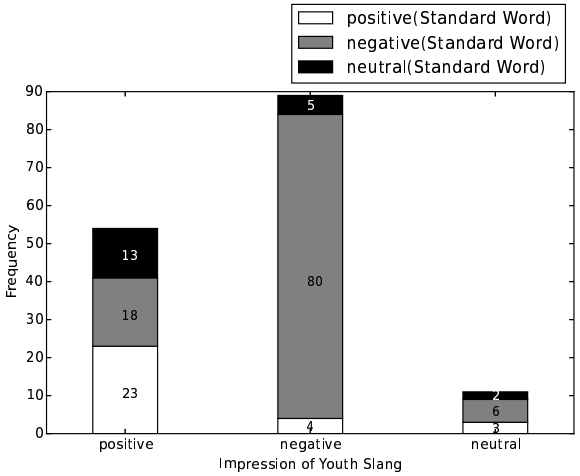

図 2: 若者言葉と標準語の positive/negative の比較

3 提案手法

本研究では，Twitter から検索キーワードとして若者言葉と感情表現について収集し，それぞれ若者言葉 Tweet コーパスと感情表現 Tweet コーパスとして構築した．この 2 つのコーパスを用いて，若者言葉を意味的かつ感性的に類似する語に変換する手法を提案する．具体的には，意味的に類似することを，文脈的に類似するかどうかにより判断するため，対象となる単語の周辺単語から学習された単語ベクトルを用いる．単語ベクトルの学習には，既存の単語ベクトル生成ツールである word2vec[3] を用いる．

[2] https://code.google.com/p/bayon/wiki/Tutorial_ja

[3] https://code.google.com/p/word2vec/

また，単語の感性的な類似性を考慮するため，感情ベクトル類似度 [11] を用いる．感情ベクトル類似度は，単語の感情を 2 次元のベクトルでとらえることで，単語間の感情的な類似性を計算したものである．肯定的/否定的および動的/静的の 2 つの軸における強度は，感情極性対応表 [8] および，感情表現辞典 [9]，日本語アプレイザル評価表現辞書 [10] に登録されている感情・評価表現との共起関係から得た値を用いる．このベクトルは，感情表現や評価表現以外の語にも付与されるため，間接的にある特定の感性を連想させるような場合にも応用できる．たとえば,「オサレ」という若者言葉は,「おしゃれ」という標準語と意味的に対応している．しかし,「オサレ」という若者言葉で表現することで皮肉や，卑下，揶揄といったネガティブな意味を含むようになる．本手法では，ポジティブな意味の「おしゃれ」だけではなく，ネガティブな意味の変換候補も得られると考える．図 3 に，提案手法の処理の流れを示す．

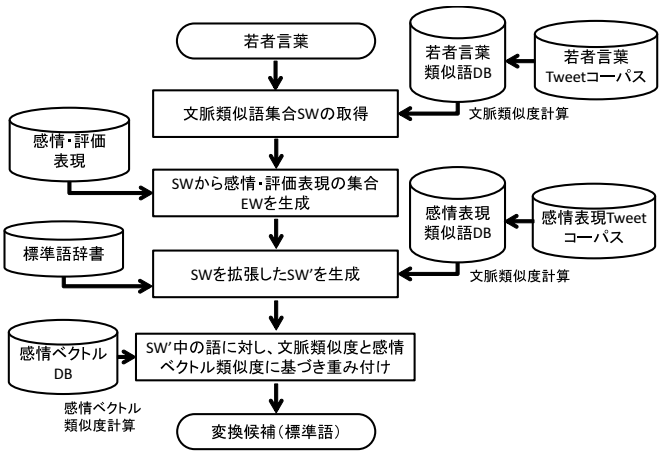

図 3: 提案手法の流れ

まず，対象となる若者言葉と文脈が類似する語を若者言葉 Tweet コーパスから検索する．つぎに，文脈類似語として得られた語のうち，感情・評価表現辞書に登録されている語と一致する語に対し，感情表現 Tweet コーパスで文脈類似語への拡張をおこなう．この拡張の際，標準語辞書に登録されている語については，その標準語と同じ意味の語を感情表現 Tweet コーパスで検索することで，より多くの表現の表現を獲得する．

獲得された標準語の候補に，類似度による重み付けをおこなう．以下，重み付けの計算方法について説明する．入力若者言葉を ys_i，若者言葉をキーとして若者言葉 Tweet コーパスから得た文脈類似語のうち，感情・評価表現に該当する語を $ew_j \in EW$ とする．また，それらの間の文脈類似度を $sim_y = csim(ys_i, ew_j)$ とする．EW 中の各標準語 sl_x に対して，後述する方法でクラスタリングした結果，類義語となった語の集合 $sy_k \in SYN$ を得る．sy_k をキーとして，感情表現 Tweet コーパスから文脈類似語の集合を得て，EW 中の語との文脈類似度の最大値 $sim_c = csim_{max}(sy_k, EW)$ および，感情ベクトル類似度の最大値 $sim_e = esim_{max}(sy_k, EW)$ を得る．重みは，これらの類似度を組み合わせて式 1 のように計算する．

$$\begin{aligned} sim_y &= csim(ys_i, ew_j) \\ sim_c &= csim_{max}(sy_k, EW) \\ sim_e &= esim_{max}(sy_k, EW) \\ W(sl_x) &= (sim_e + sim_c) \times sim_y \end{aligned} \quad (1)$$

標準語の意味クラスタリングは，日本語 WordNet，日本語語彙大系，分類語彙表，EDR 概念辞書の 4 つのシソーラス中の意味 ID を特徴としたベクトルを単語ごとに作成し，クラスタ数 k を 30,000 に設定しておこなった．

4 評価実験

4.1 若者言葉の標準語への変換実験

提案手法では Tweet コーパスをもとに若者言葉との類似語集合を取得し，そのなかから意味的かつ感性的に類似する語を複数抽出する．抽出した語を標準語への変換候補としたときに，意味的な観点と感性的な観点からの両面においての評価をおこなう．印象アンケートに用いた若者言葉 671 語を含む 1323 語を実験対象とする．

若者言葉 Tweet コーパスと感情表現 Tweet コーパスに対し MeCab による分かち書き処理を施し，若者言葉以外は基本形に変換することで，学習データを作成した．この 2 通りの学習データに対し，word2vec により単語ベクトルの学習をおこなう．学習パラメータは，単語ベクトルのサイズを 200，window サイズを 10 に設定した．

4.2 実験結果

候補として得られた標準語について，意味的に妥当であるか否かを，人手により作成した標準語リストと照らし合わせる方法と，語釈文中の単語と比較することにより評価する．具体的には，出力された標準語候

補と，人手により作成した若者言葉に対応する標準語リスト，または若者言葉の語釈文に含まれる標準語とが一致すれば正解，一致するものが無ければ不正解とみなして正解率を求める．また，感性が類似するか否かは，アンケート結果の positive/negative と，出力された標準語の上位 10 語までの感情ベクトルの和における positive/negative との一致率をみることで評価する．具体的には，極性の一致率を，一致した数を全体の数で割ることで求める．

図 4 に，2 通りの意味の一致についての評価結果を示す．図中の csim は文脈類似度のみを用いた手法，csim+esim は文脈類似度と感情ベクトル類似度を用いた手法を示している．また，interpretation of words は，語釈文中の語との比較，manual annotation は，人手による標準語候補との比較を示す．

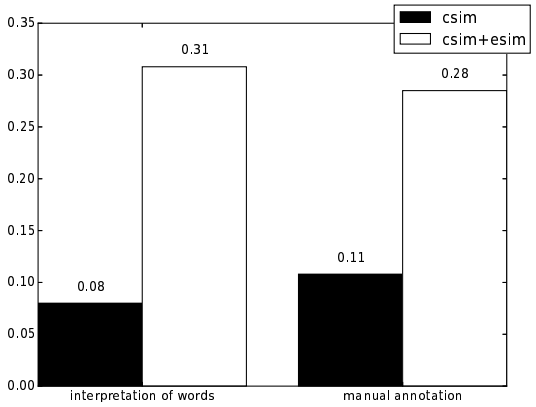

図 4: 意味による評価結果（正解率）

表 1 に，感性の一致についての評価結果を示す．意味･感性両方において提案手法 (csim+esim) の方が良い結果が得られた．また，事前調査でも positive/negative が類似しない表現が多数あったため，感性による評価において一致率が低いことは予想できた．しかし，意味による評価結果をみると，あまり高い正解率が得られなかったことから，意味による候補拡張時に，文脈類似度の値が一定以上の語のみを拡張するなどの制約が必要と考える．また，今回は positive/negative のみの評価をおこなったが，感情推定などへの応用を考えると，感情カテゴリの一致についても評価する必要がある．

表 1: 感性による評価結果（感情極性の一致率）

csim	csim+esim
0.29	0.35

5 おわりに

本稿では，ソーシャルメディア上で多用される若者言葉に着目し，意味と感性の双方において類似する標準語への変換手法について提案した．意味と感性を考慮することで，意味のみを考慮するよりも若者言葉に対して適切な標準語を得られた．しかし，提案手法では，感性を考慮することで候補数が増大するという問題がある．今後は，文脈類似度に閾値を設けることなどにより，候補を厳選する方法を考案したいと考えている．

参考文献

[1] 森信介，ニュービッググラム，“言語資源の追加:辞書かコーパスか”，情報処理学会研究報告，自然言語処理研究会報告 2014-NL-216(12), pp.1–3, 2014-05-15.

[2] 鍛冶伸裕，喜連川優，“未知語を考慮した形態素解析のための単語ラティスの効率的な生成方法”，情報処理学会研究報告. SLP, 音声言語情報処理 2013-SLP-96(12), pp.1–8, 2013-05-16.

[3] K. Matsumoto, K. Kita and F. Ren: “Emotion Estimation from Sentence Using Relation between Japanese Slangs and Emotion Expressions,” Proc. of the 26th Pacific Asia Conference on Language,Information and Computation, pp. 377–384, 2012.

[4] F. Bond, T. Baldwin, R. Fothergill and K. Uchimoto: “Japanese SemCor: A Sense-tagged Corpus of Japanese,” Proc. of the 6th International Conference of the Global WordNet, 2012.

[5] 池原悟，他： “日本語語彙大系”，岩波書店, 1997.

[6] 国立国語研究所（編） : “分類語彙表増補改訂版”，大日本図書，2004.

[7] EDR 電子化辞書，情報通信研究機構.

[8] 高村大也，乾孝司，奥村学，“スピンモデルによる単語の感情極性抽出”，情報処理学会論文誌 47(2), pp.627–637, 2006-02-15.

[9] 中村明，“感情表現辞典”, 東京堂出版, 1993.

[10] 佐野大樹： “日本語における評価表現の分類体系：アプレイザル理論をベースに”，電子情報通信学会技術研究報告. NLC, 言語理解とコミュニケーション 110(400), pp.19–24, 2011-01-20.

[11] K. Matsumoto, K. Kita and F. Ren: “Emotional Vector Distance Based Sentiment Analysis of Wakamono Kotoba,” China Communications 9(3), pp.87–98, 2012.