E-032

用例間の類似度に基づく若者言葉の感情推定手法

Estimating Emotion of *Wakamono Kotoba* Based on Similarity of Example Sentences

松本 和幸 † Kazuyuki Matsumoto　**北 研二** † Kenji Kita　**任 福継** † Fuji Ren

1. まえがき

若者言葉は，10 代から 20 代の若年層が使用する表現であり，日常会話で，主に仲間内で用いられることが特徴である [1, 2]．インターネットが幅広い世代に浸透した現在では，ネット俗語というものが，日常会話で用いられる機会が増加した．たとえば,「リア充」という言葉は，もともと Web 上でしか用いられなかった語であるが，現在では，日常会話でも用いられる．このような語も若者言葉の範疇に入れるとすれば，評判分析，意見マイニング [3] などにおけるテキスト解析処理で，若者言葉の認識が重要なタスクとなる．

若者言葉の認識における問題点を，以下に挙げる．

(a) 表記の豊富さ（曖昧さ）

(b) 新語義の存在

まず，(a) の，表記の豊富さ（曖昧さ）であるが，たとえば，若者言葉の「キモい」という表現について，同じ意味での別の表記を挙げると,「きもい」,「キモイ」,「きめぇ」,「きしょい」,「キショい」などがある．表記によってニュアンスが多少変化するが，これらはすべて同義語として扱うべきである．

また，未知語を既存のカテゴリに分類する研究が従来から行われているが [4]，そのほとんどが人名や地名，商品名などの固有名詞が主な対象である． Matsumoto[5] らは，形態素解析器 MeCab[1] を用いて 199 種類の若者言葉が含まれる 3,919 文を形態素解析し，若者言葉が 1 単語で認識される場合について品詞の統計を行った結果，表 1 のような品詞分布を得ている．形態素解析用辞書として，MeCab 標準の IPA 辞書，Naist 辞書，UniDic[6] の 3 種類で比較している．

表 1: 若者言葉の品詞分布（%）

辞書	未知語	名詞	動詞	形容詞	副詞	その他
IPAdic	61.2	26.8	5.6	5.5	0.7	0.2
UniDic	15.1	55.5	7.7	10.4	5.6	5.6
Naistdic	58.4	29.4	6.2	5.2	0.6	0.2

MeCab は，辞書に未登録の語はデフォルトでは，未知語として認識されないため，未知語は未知語として品詞タグ付けするように設定した上での統計結果である．この結果を見ても分かる通り，UniDic を除いて，一般的な辞書には，約 60%の若者言葉は，登録されていない．また，この登録されていない若者言葉について調査したところ，そのほとんどが固有名詞ではなかった．しかし，表記に曖昧性があると異表記の語は別の語として認識されることになり，文書分類タスクにおいて精度を下げる原因となる．表記を統一するために表記ゆれ解消辞書に，こうした表現の追加も必要となる．また，ある程度文字列が類似していれば，文字列の類似性を測る LCS(longest common subsequence)[7] を用いたり，規則を用いたりすることで表記ゆれが解消できると考えられるが，若者言葉に多く存在する，文字列があまり類似していない異表記に対しては解消が難しい．

次に，(b) の問題点であるが，若者言葉は，既存の語の言い換えたものであったり，省略表現も多いが，複合的な意味を持つ語や，既存の概念には属さない語も多い．こうした語のために，新語義を定義する必要があるが，その定義方法について未解決の問題が多い．我々の研究の目標は，若者言葉を含む文からの感情推定である．若者言葉を特徴として感情推定を行う際に，若者言葉を，その意味や用法，さらには感情などで分類することができれば感情推定の精度向上につながると考えられる．そこで，本稿では，若者言葉の用例に基づき，感情表現に関する情報を抽出することで，若者言葉の感情を推定する手法を提案する．

2. 関連研究

Matsumoto[8] らは，若者言葉に対して感情ベクトルを付与した若者言葉感情辞書の構築を行っている．その研究で提案されたシステムは，未知の若者言葉が入力されると，データベース中の若者言葉との表層的類似度または意味的類似度により感情ベクトルを決定して出力する．評価実験の結果，表層的特徴を用いた手法では，意味的類似度を用いた手法よりも精度が高く，正解率は約 50%程度であった．しかし，この精度は高いとは言えないため，表層的な特徴だけでは，感情の推定には不十分であると考えられる．

我々は，用例が似ている若者言葉同士は用法が似ており，意味的，さらには感情的に類似するのではないかと考え，用例間の類似性に着目する．しかし，単に用例が似ているだけで感情が似ていると判断することはできない．そこで，文の字面上の類似性だけでなく，文の感情の類似性を，単語毎の感情ベクトルを定義することで求め，それを基に用例間の類似度（距離）を計算する．

† 徳島大学大学院ソシオテクノサイエンス研究部

[1] http://mecab.sourceforge.net/

3. 提案手法

評価に用いる若者言葉は，Web 上で既に広く使用されていて，多くの人に認知されているものとする．学習データとして扱える言語資源として，松本 [9] らが構築中の若者言葉感情コーパスがある．このコーパスは，若者言葉を含む文を Web ブログ検索により収集し，書き手の感情を数名の作業者によりアノテーションしたものであり，現在，18,038 文収集済みである．

本研究で提案する手法では，従来の bag of words に基づくテキスト分類や，テキスト間の類似度計算では考慮されなかった，感情という観点からの単語間の距離尺度を用いる．機械学習に基づく感情推定において，文中の単語が感情を表すか否かにより素性選択する研究 [10] はこれまでにもあるが，単語間の感情的な関係を考慮するものは無かった．単語間の感情的な関係を何らかの距離尺度により定義することで，図 1 のように，若者言葉を含む例文集合を入力すると，文集合間の感情的な距離を計算し，最小距離を求め，最も距離が近い文から順に，その文に含まれる若者言葉の感情を出力するようなシステムを想定している．

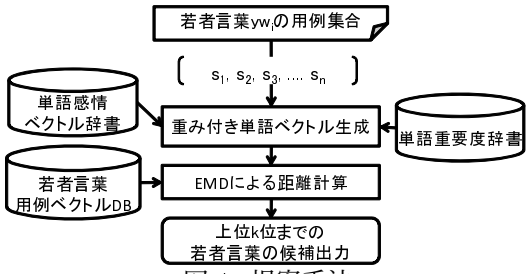

図 1: 提案手法

3.1. 若者言葉用例ベクトルデータベースの構築

ここでは，入力された若者言葉の用例との類似度を計算するための若者言葉用例ベクトルデータベースの構築手順について述べる．

1. 若者言葉感情コーパスから，若者言葉 yw_i と，その用例集合 S_i を抽出する．
2. S_i 中の，各用例 s_{ij} について，MeCab を用いて形態素解析を施し，文中の単語（基本形）の重要度を決定し，単語とその重要度を組とした単語ベクトルを作成する．ここで，機能語（助詞，助動詞など）および若者言葉に該当する部分は除去しておく．
3. 用例に付与されている感情の種類と，作成した単語ベクトルを関連付ける．

Step.1〜3 までを，コーパス中の若者言葉の全ての用例毎に行う．これらを登録したデータベースが，若者言葉用例ベクトルデータベースとなる．

文中の単語の重要度として，単語の出現頻度に基づく TF·IDF という尺度が情報検索や文書分類で一般的に用いられる．TF·IDF は，タスクに応じて改良が加えられ，拡張されたものが多数ある [11, 12]．本研究では，Lee ら [11] の定義する拡張 TF·IDF および，Okapi/BM25 の TF·IDF[12] を用いる．式 1 に，拡張 TF·IDF，式 2,3 に Okapi/BM25 の計算式を示す．$tf_{i,d}$ は，単語 i の用例集合 d における出現頻度，DF_i は，単語 i の出現する用例集合数，$|D|$ は，用例集合の総数を示す．len_a, len_d は，それぞれ，全用例集合の単語総数の平均，用例集合 d の単語総数を示す．k_1, b は，それぞれ，$tf_{i,d}$，正規化 (len_d/len_a) の重視の度合いを意味するパラメータである．本稿では，各パラメータを，$k_1 = 0.5, b = 0.5$ と設定して用いる．

$$\text{E-TF}\cdot\text{IDF}_{i,d} = \frac{(0.5 + 0.5\cdot\frac{tf_{i,d}}{\max_d tf_i})\log_{10}\frac{|D|}{\mathrm{DF}_i}}{\sum_d\left[(0.5 + 0.5\cdot\frac{tf_{i,d}}{\max_d tf_i})\log_{10}\frac{|D|}{\mathrm{DF}_i}\right]^2} \quad (1)$$

$$\mathrm{TF}_{\mathrm{BM25}}(i,d) = \frac{tf_{i,d}\cdot(k_1+1)}{tf_{i,d} + k_1\left[(1-b) + b\frac{len_d}{len_a}\right]} \quad (2)$$

$$\mathrm{IDF}_{\mathrm{BM25}}(i) = \log_2\frac{|D| - \mathrm{DF}_i + 0.5}{\mathrm{DF}_i + 0.5} \quad (3)$$

Okapi/BM25 の場合は，式 2,3 における $\mathrm{TF}_{\mathrm{BM25}}(i,d)$ と $\mathrm{IDF}_{\mathrm{BM25}}(i)$ とを掛け合わせたものを，単語の重要度とする．単に単語の重要度のみだと，感情に関連しない語（文の感情表現に貢献しない語）が多数マッチすることで，類似度が高くなってしまう問題が出る．そこで，単語の表す感情が文の表す感情と似ているほど，その単語が重要であると定義し，文の表す感情と単語の表す感情との類似度（感情ベクトル類似度）を計算する手法を提案する．そして，この類似度と，拡張 TF·IDF または Okapi/BM25 による重要度との重み付き線形和により得られた値を，単語の感情重要度として用いることにする．式 4 に，感情重要度の計算式を示す．本稿では，$\alpha = 0.3$ と設定して実験した．

$$\begin{aligned}\text{感情重要度} \;=\; & \alpha \times \text{単語重要度} \\ & + (1-\alpha)\times\text{感情ベクトル類似度}\end{aligned} \quad (4)$$

3.2. 単語感情ベクトル辞書

本研究では，Lang の感情 2 次元モデル [13] を用い，この 2 次元のベクトル表現により，感情を表す尺度を定義する．このモデルは，感情状態を，Arousal（活性）と Valence（快/不快）の 2 次元で表現するものである．例えば，活性状態で，不快であれば，〈1,-1〉 と表す．感情ベクトルの各次元の値は，-1 から 1 までの実数値をとる．本研究では，既存の辞書を組み合わせることで，単語に 2 次元のベクトル（感情ベクトル）を付与した単語感情ベクトル辞書の作成を行う．以下，単語感情ベクトル辞書の作成方法について述べる．

「感情表現辞典」[14] から，10 種類の感情に分類された 2,379 種類の単語と，高村らの感情極性対応表 [15] とを組み合わせることで，単語感情ベクトル辞書を構築する．感情極性対応表は，国語辞書に含まれる単語に，ポジティブ/ネガティブのそれぞれの強さを示す値

表 2: 感情の種類と感情ベクトルとの対応

ID	感情ベクトル	感情の種類
1	(1 , -1)	怒, 恐
2	(1 , 1)	喜, 好
3	(-1 , -1)	悲, 嫌
4	(1 , 0)	驚, 昂
5	(-1 , 1)	安
6	(0 , 0)	平静

（感情極性値）を機械的に付与した辞書である．この辞書の単語エントリ数は，日本語で 55,125 であり，語彙数としては十分であると考えられる．この辞書の単語に付与されている感情極性値を基に Valence の次元は付与できる．しかし，Arousal の次元については，別の基準を用いて付与する必要がある．

感情表現辞典に含まれる語は，感情の種類に応じて，表 2 に示す感情ベクトルに基づき，Arousal と Valence の 2 次元で表す．感情表現辞典と感情極性対応表の両方に含まれる単語は，感情表現辞典での感情の種類を基に，感情ベクトルを付与する．感情極性対応表のみに含まれる語の場合，日本語 WordNet[16] を検索し，同義語・上位語・下位語集合を得て，その単語集合中に，感情表現辞典に登録されている語が存在すれば，その語の感情ベクトルを付与する．この時，もし，この単語集合中に感情表現辞典に含まれる複数の語が存在すれば，感情ベクトルの総和をとり，その平均値により感情ベクトルを生成する．WordNet の検索によって，感情ベクトルが付与されなかった場合，感情極性の値が正か負かにより感情ベクトルを付与する．具体的には，値が正であれば,〈0,1〉, 値が負であれば,〈0,-1〉とする．

感情極性対応表において極性値の絶対値が小さいものは，あまり感情の表現に貢献しない語であると考えられる．そこで，語彙数と感情極性の信頼性を考慮して，感情極性閾値の閾値を 0.5 に設定して実験に用いる．

3.3. 感情重要度

2 つの感情ベクトル間の距離は，ユークリッド距離により定義する．ある文 s と，その文に含まれる単語 w のそれぞれの感情ベクトルを V_s, V_w と表した時，V_s, V_w 間の感情ベクトル距離 $evdist_{s,w}$ を，式 5 により求める．式中の $v_{s,i}, v_{w,i}$ は，各次元の要素を示す．n は，感情ベクトルの次元数を表す．

$$evdist_{w,s} = \sqrt{\sum_{i=1}^{n}(v_{s,i} - v_{w,i})^2} \quad (5)$$

この感情ベクトル距離 $evdist_{s,w}$ から，ある文におけるある単語の感情重要度を計算する．各単語の感情重要度 $imp_{ev}(w,s)$ は，式 6 により計算する．

$$imp_{ev}(w,s) = 1 - \frac{evdist_{w,s}}{max\_evdist} \quad (6)$$

max_evdist は，感情ベクトル距離が取り得る最大値を表す．感情ベクトルが 2 次元で，値がとりうる範囲が-1 から 1 までの実数値であれば，$max\_evdist = 2\sqrt{2}$ となる．

3.4. 感情重要度に基づく類似度計算

一般に，入力文の感情は未知であるため，文中の単語の感情重要度が求められない．そこで，文中のすべての単語について，各単語間の感情ベクトル距離を算出し，それに基づき感情重要度を求めることにする．ある語 w に対して，その他の語との感情ベクトル距離の平均値を計算したものを，w の感情重要度とする．

得られた重み（感情重要度）付き単語ベクトルと，若者言葉用例ベクトルデータベース中の重み付き単語ベクトル間の距離は，単語の感情ベクトルの類似度を平均する方法も考えられるが，単語の感情重要度を考慮するために，ヒッチコック型輸送問題の解に基づく尺度である Earth Mover's Distance[17](EMD) による距離を用いる．EMD は，類似画像検索や，音楽検索 [18]，文書検索 [19] などで応用されている距離尺度である．藤江ら [20] は，文書検索タスクにおいて，検索語と，文書との距離を EMD により求めることで，検索語が少ない場合にも柔軟な対応ができる手法を提案している．本研究では，単語間の感情的な関連性を考慮したいため，単語間の感情ベクトル距離を用いて，藤江らの研究を参考に，EMD に基づく用例間の類似度計算方法を提案する．単語感情ベクトル間の距離計算を，ヒッチコック型輸送問題に置き換えると，まず，需要地は，入力若者言葉の例文集合から生成された重み付き単語ベクトルにおける各単語と定義できる．また，供給地を，若者言葉用例ベクトルデータベースにおける用例ベクトルの各単語とする．各単語の感情重要度は，それぞれ，需要地における需要量，供給地における供給量を表す．各需要地から供給地までの距離は，単語の感情ベクトル間の距離で定義する（式 5 により計算）．EMD は，シンプソン法により最小距離を求める．単語の感情ベクトル間の距離は式 5 により求める．

4. 実験

提案手法の有効性を評価するため，構築した若者言葉用例ベクトルデータベースおよび単語感情ベクトル辞書を用いて，若者言葉の感情推定の評価実験を行った．

若者言葉と，その若者言葉を含む例文集合を入力とし，若者言葉用例ベクトルデータベース中の例文集合との距離（類似度）計算を行い，距離の近いものを候補として得る．実験は，leave-one-out-cross-validation で行う．ベースラインを，単語の重要度を，拡張 TF·IDF(E-TF·IDF)，Okapi/BM25 により計算し，単語ベクトルを生成した場合の，ベクトル空間モデル (VSM) に基づく手法とする．この際の類似度は余弦類似度を用いる．

例文数が 50 文以上の若者言葉（120 種類）を選択し，その用例 14,440 文を用い，若者言葉用例ベクトルデータベースを構築する．評価対象となる若者言葉は，データベース中の文中の全ての若者言葉である．例文数が 50 未満の語も含めると 1,011 種類（表記違いなども含む）となる．また，実験に用いた単語感情ベクトル辞

書の単語総数は，17,691 語である．

4.1. 実験結果

出力結果を，距離が小さい順（類似度の大きい順）に並べ，上位 k 位までに入力若者言葉と同じ感情タグが含まれる割合を適合率で表し，その平均を算出した．提案手法とベースラインの結果（k=1〜5) を，表 3 に示す．また，図 2，3 は，k の値を変えていった時の適合率の推移をグラフで示したものである．図中の k の値は，1〜10 までは 1 ずつ増加させ，10 以降は，15, 30, 50 としている．

表 3: 実験結果（適合率）

k	提案手法		ベースライン	
	E-TFIDF	Okapi/BM25	E-TFIDF	Okapi/BM25
1	35.7	34.3	27.4	30.2
2	36.5	34.5	26.0	29.0
3	35.0	33.7	23.6	27.8
4	34.7	32.8	22.2	27.1
5	33.5	32.2	21.2	26.4

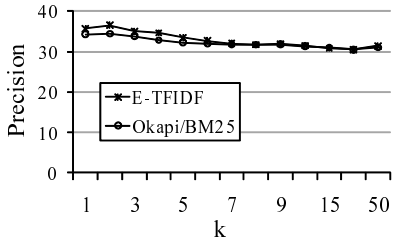

図 2: 実験結果（提案手法）

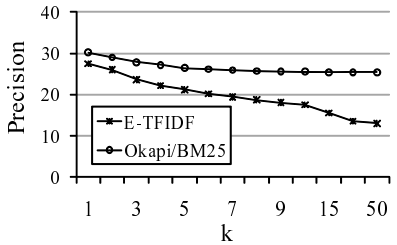

図 3: 実験結果（ベースライン）

4.2. 考察

提案手法の推定精度は，ベースラインを上回った．また，出力候補数を増やしても精度がそれほど下がらないという点は，提案手法の利点といえる．一方，提案手法では，E-TF·IDF を単語重要度として用いた方が，少し精度が高くなったが，ベースラインではこれとは逆の結果となった．これについては，感情重要度の計算における重み α や，Okapi/BM25 のパラメータの調整などでも変化する可能性がある．また，提案手法では，感情ベクトル類似度を，単語間の距離として用いたが，意味的距離との比較を行うために，同義語・類義語集合を用いた単語間距離を導入した実験も行う必要があると考えられる．

5. おわりに

本研究では，若者言葉の感情を，その語を含む例文を入力として推定する手法を提案した．評価実験の結果，単語間の距離に感情ベクトル類似度を導入し，距離尺度に EMD を用いると，余弦類似度を用いた類似度計算よりも高い精度が得られた．しかし，単語感情ベクトルの付与方法や，感情ベクトル類似度，感情重要度の計算方法には改善の余地がある．また，本稿で提案した手法は，文中の若者言葉以外の，共起する内容語を特徴として用いたが，文中において若者言葉のみが感情を表現する場合は，感情推定に失敗すると考えられる．このことから，感情を表現する既知の若者言葉は用例ベクトルに含めるようにするべきである．

今後は，単語感情ベクトル辞書の改良を行い，より適切な値を得られるよう，距離尺度の再検討を行う予定である．また，文書分類で用いられるサポートベクターマシンやその他の手法での分類結果との比較も行いたい．

参考文献

[1] 山口 仲美, “若者言葉に耳をすませば”, 講談社, 2007.

[2] 米川 明彦, “若者語を科学する”, 明治書院, 1998.

[3] 乾 孝司他, “テキスト評価分析の技術とその応用”, 情報処理, Vol.48, No.9, pp.995–1000, 2007.

[4] 後藤 和人他, “Web を用いた未知語検索キーワードのシソーラスノードへの割付け手法”, 自然言語処理, Vol.15, No.3, pp.91–113, 2008.

[5] K. Matsumoto et al. , “Analysis of Wakamono Kotoba Emotion Corpus and Its Application in Emotion Estimation,” International Journal of Advanced Intelligence, Vol.3 No.1, 2011.

[6] 伝 康晴他, “コーパス日本語学のための言語資源：形態素解析用電子化辞書の開発とその応用”, 日本語科学, Vol.22, pp.101–122, 2007.

[7] D. S. Hirschberg, “Algorithms for the Longest Common Subsequence Problem,” Journal of the ACM, Vol.24, No.4, 1977.

[8] K. Matsumoto et al., “Construction of Wakamono Kotoba Emotion Dictionary and Its Application”, Proc of CICLing2011, Vol.LNCS6608, pp.405–416, 2011.

[9] 松本 和幸他, “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 発表論文集, pp.846–849, 2011.

[10] 小川 拓貴他, “えもにゅ” における短文の感情推定について”, 情報学基礎研究会報告 2010-FI-97(2), pp.1-6, 2010.

[11] Lee J.H., et al. , “N-Grambased Indexing for Korean Text Retrieval,” Information Processing and Management, Vol.35, No.4, pp.427–441, 1999.

[12] K. Spärck Jones et al. “A Probabilistic Model of Information Retrieval: Development and Comparative Experiments (parts 1 and 2).” Information Processing and Management, Vol.36, No.6, pp.779–840, 2000.

[13] P. J. Lang, “The Emotion Probe: Studies of Motivation and Attention,” American Psychologist Vol.50, No.5, pp.372–385, 1995.

[14] 中村 明, “感情表現辞典”, 東京堂出版, 1993.

[15] 高村 大也他, “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, Vol.47, No.02, pp.627–637, 2006.

[16] F. Bond et al., “Enhancing the Japanese WordNet in The 7th Workshop on Asian Language Resources,” in conjunction with ACL-IJCNLP 2009.

[17] Y. Rubner et al. “A Metric for Distributions with Applications to Image Databases,” Proc. of the 1998 IEEE International Conference on Computer Vision, pp.59–66, 1998.

[18] 獅々堀 正幹他, “Earth Mover’s Distance を用いたハミングによる類似音楽検索手法”, 情報処理学会論文誌, Vol.48, No.1, pp.300–311, 2007.

[19] Wan, X et al., “The Earth Mover’s Distance as a Semantic Measure for Document Similarity,” Proc. of the 14th ACM International Conference on Information and Knowledge Management, pp.301–302, 2005.

[20] 藤江 悠五他, “概念ベースと Earth Mover’s Distance を用いた文書検索”, 自然言語処理, Vol.16, No.3, 2009.