E-050

# 上位概念語 n-gram を用いた感情極性推定
# Emotion classification using Hypernym n-grams

納谷 大智† 吉見 毅彦†
Daichi NAYA Takehiko YOSHIMI

## 1. はじめに

近年，人間と対話することを目的とするコミュニティロボットやチャットアプリなどの対話システムが開発されている．このようなシステムを実現するには，様々な音声言語処理技術や自然言語処理技術が必要となってくる．その中でも，人間と円滑な対話をするために，システムが人間の感情を読み取り，その感情に配慮することが重要である．人間の感情を考慮せずにシステムが発話してしまうと，対話として致命的な間違いを起こしてしまう可能性があるからである．

感情推定は，100%の精度で行うことはまだできない．なぜなら，感情が明記されていない文をコンピュータで処理をするときに，感情を正確に推定することができないからである．このため，単語集合などの，数値で表せる特徴から推定することになる．しかし，人間にさえ明示できない感情を数値化することは難しい．

これまでに，入力文の感情がポジティブであるかネガティブであるかを判別する感情極性推定手法[1]が示されている．この手法は機械学習で構築され，機械学習のための素性として単語 n-gram などが使われている．推定精度は，学習データが約 10000 文の場合で 0.57 から 0.67 であると報告されている．

本研究では，従来手法[1]の推定精度を向上させることを目的とする．具体的には，単語 n-gram を使うのではなく，感情極性推定にクラス n-gram の考え方[2]を新たに取り入れることで感情極性推定の精度が向上するかを検証する．

## 2. 従来研究

感情推定では，これまでに次のような 2 種類の研究が行われている．1 つ目は，人手で作成したパターンに基づく手法であり，もう 1 つは，機械学習を用いた手法である．

松本ら[2]は，感情生起事象文型パターンに基づいて会話文の感情を推定する手法を提案している．推定に必要なパターンを人手で登録するため，なぜその推定結果が得られたのかがわかりやすい．その反面，パターンに登録されていない未知の文章が来たときに推定が難しい．評価実験によれば，登録した感情生起事象文型パターンを用いて推定できた結果が全体の 5%程度であり，未知の文章に対応ができないことが多いとされる．

徳久ら[1]は，web から獲得した感情生起要因コーパスに基づいて感情推定を行う手法を提案している．この手法は，入力文から感情極性推定のための素性を抽出し，サポートベクターマシン（SVM）に与えることで感情極性推定を行う．機械学習を用いて推定を行うため，素性をうまく抽出できれば未知の文章が入力されたとしても精度よく推定できる可能性がある．SVM で感情極性推定手法を開発するときの素性として，単語 n-gram（$n$ = 1， 2， 3），単語の感情極性を含めた n-gram，係り受け関係にある単語が用いられている．単語の感情極性は，単語感情極性対応表[4]から得られる．

†龍谷大学 大学院 理工学研究科 Graduate School of Science and Technology, Ryukoku University

## 3. 上位概念語 n-gram を用いた感情極性推定

本研究では，機械学習を用いて感情極性推定手法を構築する．

### 3.1 上位概念語 n-gram

機械学習の素性として，上位概念語を考慮した単語 n-gram（以下，上位概念語 n-gram と呼ぶ），単語の感情極性つきの係り受け関係，単語の感情極性なしの係り受け関係の 3 種類を用いた．上位概念語 n-gram は，クラス n-gram の考え方に基づくものである．

従来手法[1]で用いられた単語 n-gram を上位概念語 n-gram に変更した理由は，意味的に同じことを表している表現も推定の材料として扱うことができるからである．単語 n-gram では入力文に出現する 2 つの単語の表記が異なると，たとえそれらが類義語であっても 2 単語の類似性を考慮できないことが問題である．

本研究の上位概念語は，シソーラスとして EDR 辞書[5]を用いることで取得できる，単語の 1 つ上位の概念に登録されている単語である．

### 3.2 上位概念語への変換

入力された単語の上位概念語への変換は，EDR 辞書の概念見出し辞書と概念体系辞書を用いて図 1 に示す手順で行う．図 1 の処理例では入力文「キノコや蓮根が入ってた」の「キノコ」が「菌類」に変換されている．

まず，入力された文の形態素解析を行う．単語（内容語）の基本形が概念見出し辞書に登録されているかを確認し，登録されていればその単語の識別番号である概念識別子を抽出する．単語が概念見出し辞書に複数登録されている場合，先頭の概念識別子を 1 つだけ抽出する．形態素解析には CaboCha [6]を用いる．

次に，抽出した概念識別子を概念体系辞書の下位概念識別子と比較し，登録があればその上位概念識別子を抽出する．概念識別子が概念体系辞書に複数登録されている場合，先頭の上位概念識別子を 1 つだけ抽出する．

最後に，抽出した上位概念識別子を概念見出し辞書で検索する．検索結果の単語に日本語概念見出しが登録されていれば，入力された単語の基本形と置き換える．検索結果に日本語概念見出しが存在しなかった場合は，該当する上位概念語がなかったと判断し，変換を行わない．

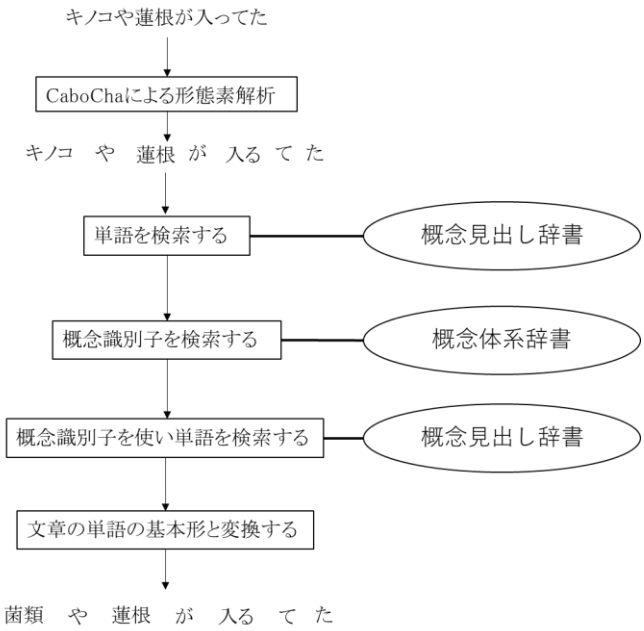


図 1　上位概念語への変換手順

## 4. コーパスの構築

本研究では，感情極性推定手法の開発にあたり，感情生起要因コーパスを用いる．このコーパスには，文の感情が生起された要因とその感情の極性を登録する．感情生起要因コーパスの構築方法は次のとおりである．

従来研究[1]と同様に，言語パターンを用いてWebテキストから自動的に感情生起要因を獲得する．言語パターンとして「X <接続表現 > <感情語>」を用いる．このパターンに Web テキスト中の文が当てはまるとき，X を感情生起要因とみなして，X と感情語の極性（ポジティブかネガティブか）を獲得する．X に，構文的な複雑さなどに関する制限は設けない．

感情語は，従来研究で感情名とされた代表的な感情語である「嬉しい」，「楽しい」，「安心」，「恐い」，「悲しい」，「残念」，「嫌」，「寂しい」，「心配」，「腹立たしい」の 10 語とする．10 語のうち「嬉しい」，「楽しい」，「安心」 の 3 語がポジティブな感情を表し，残りの 7 語がネガティブな感情を表す．また，接続表現としては，感情語が「腹立たしい」の場合は「のは」と「から」の 2 語を用い，他の感情語の場合は「のは」を用いる．例えば「理想の暮らしを考えるのは楽しい」という文が Web テキストに現れていた場合，感情生起要因 X として「理想の暮らしを考える」を，感情極性としてポジティブを獲得する．

Web テキストとして，国語研日本語ウェブコーパス[7]を用いる．このコーパスから 10 感情それぞれで 1000 文を獲得し，合計 10,000 文を本研究の感情生起要因コーパスとする．

## 5. 検証実験

提案手法が有効であるかを検証するために，次のような 3 つの実験を行う．

### 5.1 実験方法

実験 1：EDR 辞書の概念見出し辞書と概念体系辞書によって，感情生起要因コーパス中の内容語に対してどのような変換が行われるかを調査する．

実験 2：上位概念語 n-gram，単語の感情極性が付いた係り受け関係，単語の感情極性が付かない係り受け関係を素性として機械学習で提案手法を構築する．単語の感情極性は，従来研究[1]と同様に，単語感情極性対応表[4]から得る．係り受け解析には CaboCha [6]を用いる．機械学習には TinySVM [8]をオプションなしで用いる．学習データ 9000 文と評価データ 1000 文の 10 分割交差検定を行う．

提案手法と推定精度を比較するために従来手法を構築する．従来手法の構築に用いる素性は，単語-gram，単語の感情極性付きの単語 n-gram，単語の感情極性が付いた係り受け関係，単語の感情極性が付かない係り受け関係とする．従来手法に対して提案手法にどのような改善あるいは悪化が見られるかを調査する．この調査は，正解の感情極性にポジティブとネガティブの両方が含まれている場合と，正解の感情極性がポジティブだけである場合，正解の感情極性がネガティブだけである場合について行う．この調査によって，正解の感情極性がポジティブである場合とネガティブである場合で推定精度（感情極性推定の難しさ）に違いがあるかどうかを見て，もしあれば，それがどのような違いであるかを明らかにする．

実験 3：実験 2 で用いたコーパスには，ポジティブとネガティブの感情極性に 3 対 7 の偏りがある．この偏りを抑えて実験 2 と同様の検証を行う．偏りをなくすために，ネガティブの「悲しい」，「嫌」，「腹立たしい」の 3 種類とポジティブの「嬉しい」，「楽しい」，「安心」の 3 種類の計 6 感情 6000 文を用いる．

### 5.2 実験結果と考察

実験 1：コーパスの 10000 文に現れる内容語は 75456 語で，このうち上位概念語に変換された単語は 10335 語であり，変換割合は 13.70%であった．変換された単語のうち出現頻度が 10 以上のものを降順に表 1 に示す．

コーパスに出現した内容語の 13.70%しか上位概念語に変換されなかった原因の一つとして，CaboCha の解析結果と EDR 辞書の見出しの照合を，両者の間で文字列が完全に一致した場合に成功としていることが挙げられる．サ変動詞の認定単位が CaboCha と EDR 辞書で異なるため，この照合方法では失敗する．例えば「運転する」の場合，CaboCha では「運転」と「する」に分割されるが，EDR 概念見出し辞書の見出しは「運転する」だけである．このため，「運転する」と「運転」は文字列として完全一致はしない．今後，より柔軟な照合を行う必要がある．

表 1 を見ると，「一」から「宝物」，「ある」から「持ち上がる」などの明らかに不適切な変換が行われているものや，「円」から「平面図形」のように文脈によっては不適切な変換も見られた．この「円」は通貨単位であるが，図形とみなされている．不適切な変換となる原因は，3,2 節の変換で，単語が多義語である場合でも，それが使用されている文脈を考慮せずに，EDR 辞書の先頭の見出しを抽出し，機械的に 1 つに決定する処理を行っていることである．

表 1　上位概念語への変換結果

| 変換前単語 | 変換後の上位概念語 | 出現頻度 |
|---|---|---|
| なる | 移る | 86 |
| 一 | 宝物 | 80 |
| ある | 持ち上がる | 77 |
| くれる | 行う | 34 |
| いい | グッドネス | 26 |
| 気 | 香 | 18 |
| 前 | 前方 | 18 |
| 多い | 膨大だ | 16 |
| そのまま | 突然 | 16 |
| 円 | 平面図形 | 16 |
| 家 | 家系 | 16 |
| 良い | 好都合 | 15 |
| まだ | 永続的だ | 15 |
| 先生 | 人名 | 15 |
| 使う | 利用する | 14 |
| 最近 | 過去 | 13 |
| 仕事 | 仕事 | 12 |
| 今朝 | モーニング | 11 |
| もらえる | 貰う | 11 |
| 学校 | 学校 | 10 |
| いろいろ | 膨大だ | 10 |
| 意味 | 含意 | 10 |

他の変換については，おおむね妥当な変換が行われている．「多い」と「いろいろ」が共に「膨大だ」に変換され，これらが同一視されている．「学校」から「学校」のように変換前後の単語が同一のものがあるが，変換前は「機関としての学校」であり，変換後は「組織としての学校」であるため，正しく変換が行われている．

実験 2：表 2 に， 10 感情 10000 文を学習・評価データとして構築した提案手法と従来手法による推定の結果を示す．表の数値は，10 分割交差検定での推定精度 Accuracy の平均と標準偏差である．

実験 2 で用いたコーパスでは 70%がネガティブであるため，ベースラインの推定精度を 70%とする．提案手法と従来手法とベースラインの間でウェルチ法による T 検定を，ボンフェローニ補正（比較回数 3 回）による有意水準 0.02（= 0.05 / 3）で行った．

その結果，両手法の平均の差は有意であった（$t(13) = 17.22$，$p < 0.02$）．また，従来手法とベースラインの間に有意差が認められた（$t(9) = 20.16$，$p < 0.02$）が，提案手法とベースラインの間に有意差は認められなかった（$t(9) = 1.14$，$p = 0.28$）．したがって，10 感情を学習・評価データとした場合，提案手法の推定精度は，ベースラインと同程度であり，従来手法よりも高いと言える．

表 2　10 感情での推定精度の比較

| | 提案手法 | 従来手法 |
|---|---|---|
| 検定回数 | 10 | 10 |
| 平均 | 69.48% | 52.87% |
| 標準偏差 | 1.44 | 7.24 |

実験 3：表 3 に， 6 感情 6000 文を学習・評価データとして構築した提案手法と従来手法による 10 分割交差検定の結果を示す．実験 3 で用いたコーパスでは感情極性に偏りがないため，ベースラインの推定精度を 50%とする．提案手法と従来手法とベースラインの間でウェルチ法による T 検定を，ボンフェローニ補正による有意水準 0.02 で行った．

その結果，両手法の平均の差は有意であった（$t(17) = 4.82$，$p < 0.02$）．また，従来手法とベースラインの間に有意差が認められ（$t(9) = 7.51$，$p < 0.02$），提案手法とベースラインの間にも有意差が認められた（$t(9) = 12.08$，$p < 0.02$）．したがって，偏りのない 6 感情の場合も，提案手法の推定精度は，従来手法とベースラインよりも高いと言える．

表 3　6 感情での推定精度の比較

| | 提案手法 | 従来手法 |
|---|---|---|
| 検定回数 | 10 | 10 |
| 平均 | 59.47% | 54.65% |
| 標準偏差 | 2.48 | 1.96 |

表 4 に，従来手法に対して提案手法で改善した文と悪化した文の分布を示す．表 4 は，正解の感情極性にポジティブとネガティブの両方が含まれている場合の分布である．

表 4 から，従来手法に対する提案手法の悪化が推定結果の 20.27%（= 1216 / 6000）を占めるが，他方，推定結果の 25.07%（= 1504 / 6000）で改善が見られる．

表 4　従来手法に対する提案手法の改善と悪化：正解の極性がポジティブとネガティブの混在の場合

| 提案手法＼従来手法 | 正 | 誤 |
|---|---|---|
| 正 | 2064 | 1504 |
| 誤 | 1216 | 1216 |

表 5 に，正解の感情極性がポジティブだけの場合の，提案手法の改善と悪化の分布を示す．また，表 6 に，正解の感情極性がネガティブだけの場合の分布を示す．

表 5　従来手法に対する提案手法の改善と悪化：正解の極性がポジティブの場合

| 提案手法＼従来手法 | 正 | 誤 |
|---|---|---|
| 正 | 714 | 799 |
| 誤 | 688 | 799 |

表 6　従来手法に対する提案手法の改善と悪化：正解の極性がネガティブの場合

| 提案手法＼従来手法 | 正 | 誤 |
|---|---|---|
| 正 | 1350 | 705 |
| 誤 | 528 | 417 |

まず，正解の感情極性がポジティブであるかネガティブであるかによって，提案手法の推定精度が異なるかを調べる．推定精度は，ポジティブの場合は表 5 より 50.04%（= (714 + 799) / 3000）であり，ネガティブの場合は表 6 より 68.50%（= (1359 + 705) / 3000）である．したがって，正解の感情極性がネガティブの場合のほうが，提案した素性の有効性がより高いと言える．

次に，正解の感情極性がポジティブであるかネガティブであるかによって，従来手法で誤推定された文のうち提案手法で正しく推定された文の割合が異なるかを調べる．この割合は，ポジティブの場合は表 5 より 50.00%（= 799 / (799+799)）であり，ネガティブの場合は表 6 より 62.80%（= 705 / (705+417)）である．したがって，このことからも，正解の感情極性がネガティブの場合のほうが，提案した素性の有効性がより高いことが示された．

## 6. おわりに

本研究では，従来の感情極性推定手法と異なり，クラス n-gram の考え方に基づいて上位概念語 n-gram を新たに取り入れることで感情極性の推定精度が向上するかを検証した．従来手法では，感情極性推定のための素性として単語 n-gram が用いられている．しかし，単語 n-gram には，入力文に出現する単語間の類似性を考慮できないという問題がある．

検証の結果，シソーラスとして EDR 辞書を用いて変換した時，単語 n-gram に基づく従来手法よりも提案手法の推定精度が有意に高いことを確認した．

今後の課題として，従来手法に比べて提案手法で悪化する文を減らすために新たな素性を検討することが挙げられる．また，シーラスの検索方法や使用するシソーラスを見直し，適切な上位概念語が取得できるようにする必要もある．

### 参考文献

[1] 徳久良子，乾健太郎，松本裕治，"Web から獲得した感情生起要因コーパスに基づく感情推定"，情報処理学会論文誌，Vol. 50，No. 4，pp. 1365-1374 (2009)
[2] P. F. Brown, V. J. D. Pietra, P. V. deSouza, J. C. Lai and R. L. Mercer, "Class-Based n-gram Models of Natural Language", Computational Linguistics, Vol. 18, No. 4, pp. 467-480 (1992)
[3] 松本和幸，三品賢一，任福継，黒岩眞吾，"感情生起事象文型パターンに基づいた会話文からの感情推定手法"，自然言語処理，Vol. 14，No. 3，pp. 239-271 (2007)
[4] 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，Vol. 47，No. 2，pp. 627-637 (2006) http://www.lr.pi.titech.ac.jp/~takamura/pndic_ja.html
[5] 荻野孝野，仲尾由雄，小笠原あゆみ，長澤陽子，"日本電子化辞書研究所における概念体系"，情報処理学会，研究報告情報学基礎，pp. 27-34 (1993) https://www2.nict.go.jp/ipp/EDR/JPN/Intro.html
[6] 工藤拓，松本裕治，"チャンキングの段階適用による日本語係り受け解析"，情報処理学会論文誌，Vol. 43， No. 6，pp. 1834-1842 (2002) http://taku910.github.io/cabocha/
[7] 浅原正幸，河原一哉，大場寧子，前川喜久雄，"『国語研日本語ウェブコーパス』とその検索系『梵天』"，情報処理学会論文誌，Vol. 59，No. 2，pp. 299-305 (2018) https://masayu-a.github.io/NWJC/
[8] T. Kudo, "TinySVM: Support Vector Machines" (2002) http://www.chasen.org/~taku/software/TinySVM/