概要

言語の意味理解の一つとして，言語表現から書き手や登場人物の情緒を推定する技術に期待が寄せられている．情緒推定の手法は，パターン辞書を用いる手法や機械学習を用いる手法がある．これらの手法は，テキストに例えば「嬉しい」など直接的に情緒が表現されなくても，情緒生起原因を根拠として，情緒推定を行う．情緒推定技術を高める上で，情緒生起原因を表す言語表現の収集は基礎的で重要な課題である．そこで本研究では，Web コーパスから，2 種類の方式，すなわち因果表現に基づく収集方式，および，共起頻度に基づく収集方式で情緒生起原因を収集することを試みる．

一つ目の方式は，第一に，次の条件で文を収集する．接続表現「ので」となる因果表現文であること，主節が情緒の直接表現であること，文の末尾が動詞，補助動詞，もしくは助動詞が終助詞であること，述語のモダリティが順接的であるもの (例えば否定文)，かつ，句点で終了することとする．第二に，従属節から,「格要素の名詞」と「述語の動詞」，および,「それらの 2 つ組」をそれぞれ抽出する．最後に，主節の評価極性との共起に注目しながら「名詞」,「動詞」，および「2 つ組」に対して，頻度を求める．二つ目の方式は，第一に，次の条件で文を収集する．Web コーパスから「良い」（Positive の明確な形容詞），および「悪い」（Negative の明確な形容詞）が出現する文数をそれぞれ求める．第二に，因果表現文で収集した 2 つ組を含む文を収集する．その中で,「良い」と共起する文数，および「悪い」と共起する文数を求める．最後に，以上の文数を用いて評価極性値 *SO-Score* を算出する．

本研究で収集する情緒生起原因は，名詞と動詞の「2 つ組」に着目し，Kawahara らの 5 億文 Web コーパスから収集を試みる．因果表現に基づく収集では，獲得の手がかりとなる感情表現に小林らの直接表現辞書から人手で抽出した 374 語の感情表現を使用する．テキストと感情表現を照合して抽出し，従属節を形態素解析と係り受け解析によって「名詞」,「動詞」，およびそれらの「2 つ組」を抽出して，頻度を求める．共起頻度に基づく収集では，5 億文のテキストと「良い」，および「悪い」を照合して，照合した文数を求める．さらに，5 億文のテキストと「2 つ組」(Positive 傾向である名詞「写真」,「店」および，Negative 傾向である名詞「家」,「車」についての「動詞」を組み合わせた 2 つ組)

を照合する．再び抽出文と「良い」，および「悪い」を照合して，照合した文数を求める．求めた文数を式変形した *SO-Score* の式に代入して算出する．

収集した結果，因果表現文は 12,060 文収集し，10,333 の「2 つ組」を得た．また，共起頻度に基づいた文は 284,720 文収集し，その中から 205 の「2 つ組」を得た．これらを評価したところ，因果表現文からの収集は，日本語語彙大系の日本語表現パターン数に比べて数が不足している．また共起頻度に基づく収集は，網羅性が高いが，Negitive 傾向の名詞に対する評価極性の信頼性確保が課題として残った．これには多くの 2 つ組の収集を行いながら対策を考える必要がある．

2 つの方式で情緒生起原因を収集することができたが，情緒推定技術の重要な資源確保，および，言語の意味理解に関する知見獲得のために，今後も収集と考察を続ける必要があると考えている．

目 次

図 目 次

表 目 次

第1章　はじめに

言語の意味理解の一つとして，言語表現から書き手や話者，登場人物の情緒を推定する技術に期待されている．情緒を推定する技術は，テキストマイニングへの応用に可能性があると考えられている．例えば，インターネット上の掲示板やブログなどに蓄積されたテキストデータから情緒を推定することで，商品開発や社会事情に対する大衆の気持ちを知るといったことが挙げられる．

情緒を推定の手法には，パターン辞書を用いる手法 [1] や機械学習を用いる手法 [2] がある．これらによると，テキストに例えば「嬉しい」など直接的に情緒が表現されなくても，情緒生起の原因 (本論では情緒生起原因と呼ぶ) を根拠として情緒推定が行われる．パターン辞書を用いる手法では，情緒生起の原因と語義の関係を定める基準をパターン辞書に設けることで，入力文から照合された日本語パターンから情緒を推定することができる．また，機械学習を用いた手法では，あらかじめ人の感情生起に関する用例文を原因表現と共にコーパスとして蓄えておき，解析の際，入力文から原因を検出することで，感情を解析する．

情緒推定技術を高める上で，情緒生起原因を表す言語表現の収集は基礎的で重要な課題である．情緒生起原因によって，情緒推定における精度の向上が左右される．パターン辞書に情緒生起の原因と語義の関係を定める方法では，その関係を示す定義を増加させることで精度が向上する傾向がある．入力文に対する定義設計が確立しているならば，人が感情を認識するメカニズムに近い情緒推定が行えると考えられる．しかし，言語知識を全て辞書に加えることはコストの面で非効率である短所がある．一方，機械学習による情緒推定は，用例文を増加することで精度が増加する．機械的に言語知識を加えることが可能であるので効率的だが，膨張した知識をコーパスから読み込むことで，情緒推定による実行時間がかかってしまう短所がある．情緒生起原因の形式には，上記に挙げたように様々であるが，いずれもあらゆるテキストに対して情緒を推定するためにも言語資源の確保が必要である．収集した言語資源を分析することによって，言語の意味理解に関する見解が生まれ，新たな情緒推定の手法を見い出せると考える．

そこで本研究では，過去の機械的な手法を用いて，情緒生起原因を収集することを試

みる．ここで，情緒生起原因は，名詞と動詞の「2つ組」に着目している．

第2章　関連研究

本章では，情緒推定の方法をパターンベースの手法と機械学習による手法を第 2.1 節と第 2.2 節で述べ，第 2.2 節では機械学習によって情緒生起原因を収集する手法を紹介している．第 2.3 節は，情緒生起原因を収集する見地とは異なるが，評価極性を分類する手法について紹介する．

2.1　情緒属性付き結合価パターン辞書を用いた情緒推定

情緒属性付き結合価パターン辞書を用いた情緒推定で使用されている情緒原因は，文献 [3] で 125 種類が示されている.《喜び》,《悲しみ》,《好ましい》,《嫌だ》,《驚き》,《期待》,《恐れ》,《怒り》の基本情緒 8 種類の情緒名に対して，情緒原因の特徴が階層的に定義されている．階層構造を見ると，下位の特徴は，上位の特徴を継承したより具体的な特徴となっている．図 2.1 に《期待》の一例を示す.

また田中ら [1] は，日本語語彙大系に,「情緒属性」として，判断条件，情緒名，情緒原因，および，情緒対象を追加し，情緒推定用結合価パターン辞書を作成することで，パターンベースでの情緒推定の手法を示した.

情緒推定の方法は，もし，入力文と結合価パターンがマッチし，意味属性制約を充足し，かつ，判断条件が成立するならば，対応する情緒属性を出力するという手法である.

日本語語彙大系は，日本語の用言約 6,000 語について表現構造を結合価パターン 14,819 件にまとめたものである．この結合価パターンに対して，情緒属性が 11,712 セットが付与された．図 2.2 に付与されたレコードの一例を示す.

図 2.2 では，例えば「太郎がプレゼントを彼女に買う」という文であれば，パターンとマッチする．そして「彼女」が「プレゼント」に対して好ましいという判断条件が成立した場合,「彼女」は「プレゼント」に対して情緒原因「獲得」による「喜び」の情緒が推定される.

情緒属性付き結合価パターン辞書の開発後も，情緒属性の付与の補修が行われているが，人手によって情緒成立の定義の設計を行うため，不明確なパターンも存在する．一例を図 2.3 に示す．例えば，図 2.3 のパターンにマッチする文は「ビスケットが湿る」,

（期待：好都合なことが起こることを予測した
　生理的（内的な治癒，外的な治癒）
　心理的（
　　目標実現（
　　情報収集（成行き，終了直前）
　　計画（成算）
　　対人関係（
　　仲間意識（同意，同感，協力，仲直り）
　　優劣関係（優越，賞賛，服従，厚遇，保護））
　その他）

図 2.1: 《期待》の情緒原因の特徴フレーム

◆結合価パターン：$N1$ が $N2$ を $N3$ に $N4$ で 買う
◆意味属性制約：$N1$(3 主体)$N2$(*)$N3$(3 主体)$N4$(2585 数量)
◆情緒原因：〈獲得〉
◆判断条件：目標実現・近 ($N3$,$N2$)
◆情緒原因：〈獲得〉
◆情緒主：$N3$　◆情緒対象：$N2$　◆情緒名：《喜び》

図 2.2: 情緒属性付き結合価パターン辞書のレコードの一例

あるいは「おしぼりが湿る」という文である．ここで，情緒属性の付与を行うとき，$N1$ に代入される名詞 (「ビスケット」あるいは「おしぼり」) によって付与する情緒名を定めることができない．ビスケットが湿ることは直感的に好ましくないと考えられる．おしぼりが湿ることは直感的に好ましいと考えられる．

このように，マッチしたパターンでも，単語に含まれる情報を知識ベースから参照しなければ情緒を推定することが難しい事例が問題となっている．単語の知識ベースを人手によって加えることは，あまりにコストがかかる．そこで，機械学習のアプローチから足りない語彙知識を追加する必要があると考えられる．本研究では，名詞と動詞を組合せた「2 つ組」に着目し，機械的な収集を試みる．

◆結合価パターン：$N1$ が 湿る
◆意味属性制約：$N1$(2 具体 2610 場

図 2.3: 付与されていないレコードの一例

2.2 Webから獲得する感情生起要因コーパスと感情極性の推定

文献 [2] によれば，人が感情を生起する要因を Web コーパスから自動獲得した．そこでは，Kawahara らの Web コーパス [5]（5 億文）を使用した．これにより,《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》，および,〈neutral〉（ユーザ発話が感情的な意味を持たないことを表す）の 11 種類の感情を推定した．

Positive と Negative の事態を抽出するためには，寺村の定義 [6] を参考として，次の基準を用いた．

X = 感情主，Y = 対象，Z = 当該語のとき,「X は Y を Z」「X は Y に Z」「X は Y が Z」のいずれかが表現できれば，Z は感情表現である．

この定義に従い，小林らの直接表現辞書 [7] から 349 語の感情表現を得ることができた．表 2.1 に抽出した感情ごとの感情表現の数と例を示す．

そして，図 2.4 に示す言語パターンを用いることで Web コーパスから自動的に感情生起要因 (本論の情緒生起要因と同等の意味をもつ．この文献では感情生起に関する用例文を示す．) を獲得した．接続表現には 8 種類（ので，から，ため，て，のは，のが，ことは，ことは，ことが）を用いた．たとえば,「突然雨が降り出した のは がっかりだ」という文からは,〈がっかり〉が生起する要因として｛突然雨が降り出した｝を獲得する．

表 2.1: 感情生起要因コーパスの規模と例

感情極性	10 感情	感情表現 (349 語)	
		計	例
Positive	嬉しい	90	嬉しい，狂喜，喜ぶ，歓ぶ
	楽しい	7	楽しい，楽しむ，楽しめる
	安心	5	安心，ほっと
Negative	恐い	22	恐い，怖い，恐ろしい
	悲しい	21	悲しい，哀しい，悲しむ
	残念	15	がっかり，がっくり
	嫌	109	嫌，嫌がる，嫌い
	寂しい	15	寂しい，淋しい，わびしい
	心配	17	不安，心配，気がかり
	腹立たしい	48	腹立たしい，腹立つ，立腹

図 2.4: 感情生起要因を獲得するための言語パターン

感情生起要因を用いた感情推定の手法として，単語，単語自体の感情極性，係り受けの情報を特徴量として文の感情極性 (本論では評価極性とも呼ぶ) を推定した．図 2.5 に「福祉の費用の負担が増えてしまう」という文に対して，単語の感情極性つきで記述した例を示す.「福祉」は Positive,「費用，負担」は Negative の感情極性を持つとする．図 2.5 のラティスに対して，例えば 3-gram の列を展開すると,「福祉の費用，Pos. の費用，福祉の Neg.，Pos. の Neg.，の費用の，の Neg. の，…」などが得られる．これらを素性として SVM(学習には TinySVM: http://chasen.org/˜taku/software/TinySVM/) で学習して感情極性推定モデルを構築する．

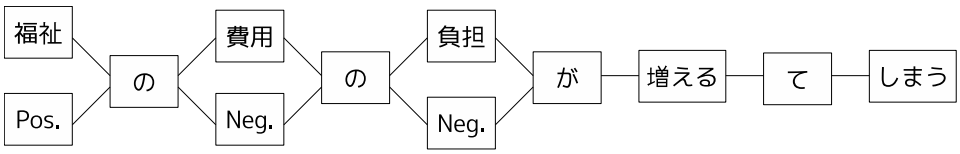

図 2.5: 感情極性つきのラティスの例

2.3 Turney らの評価極性の分類

情緒生起要因を収集する見地とは異なるが，評価極性を分類する処理について議論されている研究がなされている．その中でも，コーパスから得られる共起情報を用いて語句の評価極性値 (評価極性の傾向を示す値) を判定する手法を Turney[8] は考えた．国語辞書などのエントリ情報を用いないため，見出し語単位やエントリ単位，複数語からなる句に対しても評価極性値を判定することができる．主に，好評表現 (Positive 表現) と不評表現 (Negative 表現) の出現比率を用い，好評表現の方が多い場合は好評，逆なら不評とした．Turney は *SO-Score* を算出することで，これを示した．

ある評価表現tの極性評価値$So\text{-}Score(t)$は以下の式より算出する．ここでPMI(Pointwise Mutual Information)とは，2つの語句間の共起を図る尺度を表す．

$$SO\text{-}Score(t) = PMI(t, \text{“}Excellent\text{”}) - PMI(t, \text{“}Poor\text{”}) \tag{2.1}$$

$$PMI(a, b) = \log_2 \frac{p(a, b)}{p(a) * p(b)} \tag{2.2}$$

p(a,b)はコーパス内において単語aと単語bが同一文で共起する確率，p(x)は単語xを含む文がコーパス内で出現する確率を表している．

$So\text{-}Score(t)$で評価表現tが“Excellent”と多く共起しやすければ，正に大きい値をとり，“Poor”と多く共起しやすければ逆に負に大きい値をとる．確率が0となる語句に関しては，logに0が入ってしまうのを避けるために，Turneyらは出現頻度に0.01を足している．また，$So\text{-}Score$を算出する際に，好評文と不評文での出現頻度が共に4より小さい評価表現は有効なデータとして扱わないこととしている．

第3章　因果表現文からの収集

本章では，第2.2節で述べたWebコーパスから情緒生起原因を獲得する手法を用いて，実際に情緒生起原因を収集を行う様子を述べる．

3.1　収集方法

情緒生起原因を以下の手順で収集する．

手順 1-1 Webコーパスから次の条件で文を収集する．

- 接続表現「ので」となる因果表現文であること．
- 主節が情緒の直接表現であること．
- 文の末尾が終止形であること．
- 句点で終了すること．
- 否定文ではないこと．

手順 1-2 従属節から,「格要素の名詞」,「述語の動詞」，および,「それらの2つ組」をそれぞれ抽出する．(「ので」より前の部分を従属節とする．)

手順 1-3 主節の評価極性との共起に注目しながら,「名詞」,「動詞」，および「2つ組」に対して，頻度を求める．

3.2　収集器の実装

収集器の実装図を図 3.1 に示す．また，手順 1-1 の収集の様子を図 3.2 に，手順 1-2 の様子を図 3.3 に具体的に示す．

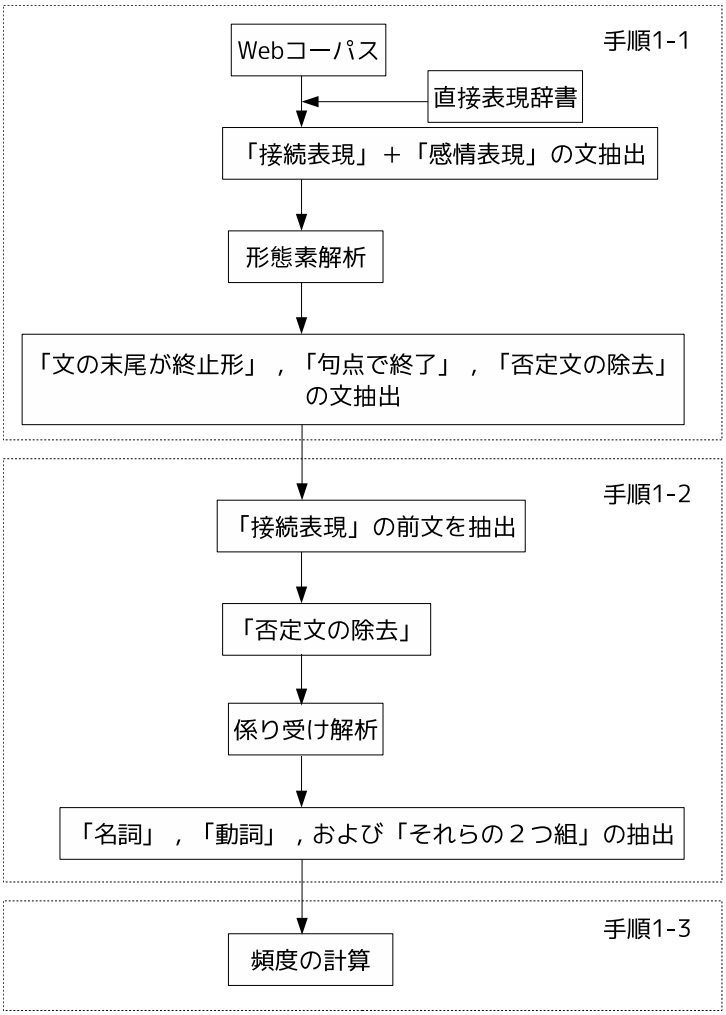

図 3.1: 収集器の実装

図 3.2 の説明

「接続表現」＋「感情表現」の文を抽出した後の形態素解析の処理結果から,「文の末尾が終止形」,「句点で終了」,「否定文の除去」の文抽出を行う工程である．1 行目に文の ID 番号，2 行目に入力文，3 行目以降が処理結果となっている．また，括弧内の番号は品詞活用形コードを示している．

7 行目の接続表現「ので」を手がかりに，次行の 8 行目に感情表現を正規表現により確認する．もし，8 行目に読点が確認された場合は，読点の次行に感情表現が確認できれば，成立することとしている．

次に感情表現を手がかりに次行の 9 行目に対して，品詞コードから「終止形」が確認された場合，成立する．成立しない場合でも，10 行目に助動詞，補助動詞，あるいは，付属語が確認された場合は，10 行目の活用形によって「終止形」が確認されると成立する．また，主節の意味が反転してしまうことを防ぐために，打消の助動詞を確認すると不成立とする．

最後に終止形を確認した次行が「句点」となった場合に抽出文として最終的な成立であると判断する．

```
1  INPUT=id001
2  風船が割れたので悲しいです。
3   /風船 (1100)
4   +が (7410)
5   /割れ (2213)
6   +た (7217)
7   +ので (7640)
8   /悲しい (3106)
9   +です (7246)
10  +。(0110)
11   /nil
```

図 3.2: 形態素解析実行結果

図 3.3 の説明

手順 1-1 で抽出した文の従属節にあたる「接続表現」の前文に対して係り受け解析を行う工程である．∗で始まる行に対し，文節の開始位置の情報が付与されている．∗の後に，文節番号，係り先番号（係り先が無い場合は −1），係り関係のスコアが付与される．

まず，終わりの文節情報を確認し,「動詞」を抽出する．もし,「動詞」+「助詞，接続助詞」+「動詞」のような複合動詞が確認された場合は，統合させて抽出を行う．

次に抽出を行った文節に対して係り先番号が付与された文節を探索する．探索された文節の中から「名詞」を抽出する．もし,「名詞」+「名詞」のような連語が確認された場合は，統合させて抽出を行う.「名詞」を含む係り先番号が付与された文節は，複数存在した場合は，全て抽出を行う．

図では 1 行目に，4 行目から始まる文節に対して係り先情報が確認されるので,「風船が」は「割れた」に対して係り受け構造を示している．最後に,「名詞」と「動詞」の 2 つ組として抽出する.「名詞」が複数存在する場合は，場合分けを行い複数の「2 つ組」として抽出を行う．

```
1  * 0 1D 0/1 0.00000000
2  風船 フウセン 名詞-一般
3  が ガ 助詞-格助詞-一般
4  * 1 -1O 0/1 0.00000000
5  割れ ワレ 動詞-自立
6  た タ 助動詞
7  EOS
```

図 3.3: 係り受け解析実行結果

因果表現文 風船が割れた/ので/悲しい.

名詞 = 風船，動詞 = 割れた，2 つ組 = (風船，割れた)，評価極性 = Neg.

3.3 収集結果

手順 1-1 の結果，評価極性が Positive の因果表現文は 10,066 文，同じく Negative の文は 1,994 文となった．手順 1-3 の結果は表 3.1 に示す．表 3.2 には「名詞」，表 3.3 には「動詞」，表 3.4 には「2 つ組」の具体例を示す．

表 3.1: 収集結果

分布	異なり数	延べ数	頻度 1 の数
名詞	5,748	2,075,028	3,512
動詞	1,220	2,211,860	609
2 つ組	10,333	816,307	9,423

※単位は名詞と動詞が語数，2 つ組が組数

表 3.2: 名詞の頻度

順位	頻度	Pos.	Neg.	名詞
1	361	291	70	人
2	185	165	20	自分
3	140	140	0	日本語
4	102	93	9	手
5	101	82	19	他
6	90	89	1	スタッフ
7	84	76	8	店
7	84	73	11	感じ
9	77	65	12	気
10	75	61	14	本
11	73	59	14	目
12	71	60	11	内容
13	63	53	10	作品
14	62	54	8	日本
15	60	50	10	場所
16	58	46	12	車
16	58	42	16	顔
18	57	54	3	写真
19	56	40	16	家
20	54	47	7	情報
21	53	47	6	先生
22	51	37	14	状態
		- - -		
2,237	1	1	0	１１０番

表 3.3: 動詞の頻度

順位	頻度	Pos.	Neg.	動詞
1	1,813	1,385	428	ある
2	625	615	10	できる
3	447	271	176	なる
4	410	303	107	いる
5	255	169	86	する
6	244	243	1	出来る
7	220	140	80	違う
8	183	153	30	思う
9	178	148	30	出る
10	159	123	36	いう
11	130	116	14	見える
12	122	119	3	わかる
13	120	118	2	使える
14	92	82	10	入る
15	82	70	12	変わる
16	80	57	23	来る
		- - -		
672	1	1	0	ＵＰする

表 3.4: 2 つ組の頻度

順位	頻度	Pos.	Neg.	(〈名詞〉, 〈動詞〉)
1	79	55	24	(人, いる)
2	24	20	4	(感じ, する)
3	23	22	1	(手, 入る)
4	20	19	1	(人, いた)
5	19	19	0	(スタッフ, いる)
6	16	16	0	(手, 入った)
7	15	5	10	(部分, ある)
7	15	15	0	(日本語, 通じる)
9	13	13	0	(看板, 出ている)
9	13	6	7	(楽しみ, していた)
11	12	11	1	(実績, ある)
11	12	11	1	(興味, ある)
11	12	12	0	(気, なっていた)
11	12	12	0	(看板, ある)
15	11	11	0	(買取, 担当する)
15	11	11	0	(人, ある)
15	11	11	0	(興味, あった)
			- - -	
911	1	1	0	(１つ, できる)

3.4 評価

日本語語彙大系 [9] によると，日本語の名詞は約 400,000 語，日本語の基本的な用言は約 6,000 語であり，日本語の用言の表現構造は，約 14,800 パターンである．一般に，複合動詞も追加する必要があるので，さらに多くの語と語義が必要と言われている．しかし，収集結果によると名詞は 5,748 しかなく，日本語語彙大系に収録された名詞に対して 80 分の 1 しかなく圧倒的に不足している．そのうち頻度 1 の数は 3,512 で半数以上であるため，信頼性の確保ができない．同様に動詞 (見出し語) の数も 1,220 しかなく，基本的とされる 5 分の 1 しかない．頻度 1 のものは 609 も存在しており，そのままでは信頼性が確保できない．さらに，名詞と動詞を組み合わせた 2 つ組については，異なり数が 10,333 件であり，同じくパターン数と比べて不足している．特に頻度 1 のものは 9,423 も存在しており，因果の後ろ盾がある文ではあるが，このままでは統計的見解をすることができない．日本語語彙大系のパターン辞書を用いた情緒推定に対して，カバー率を同程度にするためにも，近似解としてでも規模の拡大が必要であると考える．

第4章　共起頻度に基づく収集

本章では，第 2.3 節で述べた Turney らの評価極性を分類する手法を用いて，第 3 章で収集した 2 つ組に対して，評価極性値を算出する様子を述べる．

4.1　収集方法

3.2 節の 2 つ組について，*SO-Score* を算出する．

手順 2-1 Web コーパスから,「良い」(Pos.) の出現する文数，および「悪い」(Neg.) の出現する文数をそれぞれ求める．

手順 2-2 Web コーパスから算出目標の 2 つ組を含む文を抽出する．その中で,「良い」と共起する文数，および「悪い」と共起する文数をそれぞれ求める．

手順 2-3 以上の文数を用いて 2 つ組の *SO-Score* を算出する．また，4.1 式は 4.2 式より 4.3 式のように式変形できる．これにより，手順 2-1，手順 2-2 で求めた 4 つの出現文数を用いて算出する．

$$SO\text{-}Score(t) = PMI(t, \text{“良い”}) - PMI(t, \text{“悪い”}) \tag{4.1}$$

$$PMI(a, b) = \log_2 \frac{p(a, b)}{p(a)p(b)} \tag{4.2}$$

$$SO\text{-}Score(t) = \log_2 \frac{N(\text{悪い}) * N(t, \text{“良い”})}{N(\text{良い}) * N(t, \text{“悪い”})} \tag{4.3}$$

ただし，手順 2-2 での共起頻度が低い場合，統計的な信頼性が得られないので，*SO-Score* は算出できない．本研究では，文献 [8] にならい,「良い」,「悪い」の共起する文数の和が 5 以上の 2 つ組を扱うことにする．

4.2 収集器の実装

収集器の実装図を図 4.1 に示す.

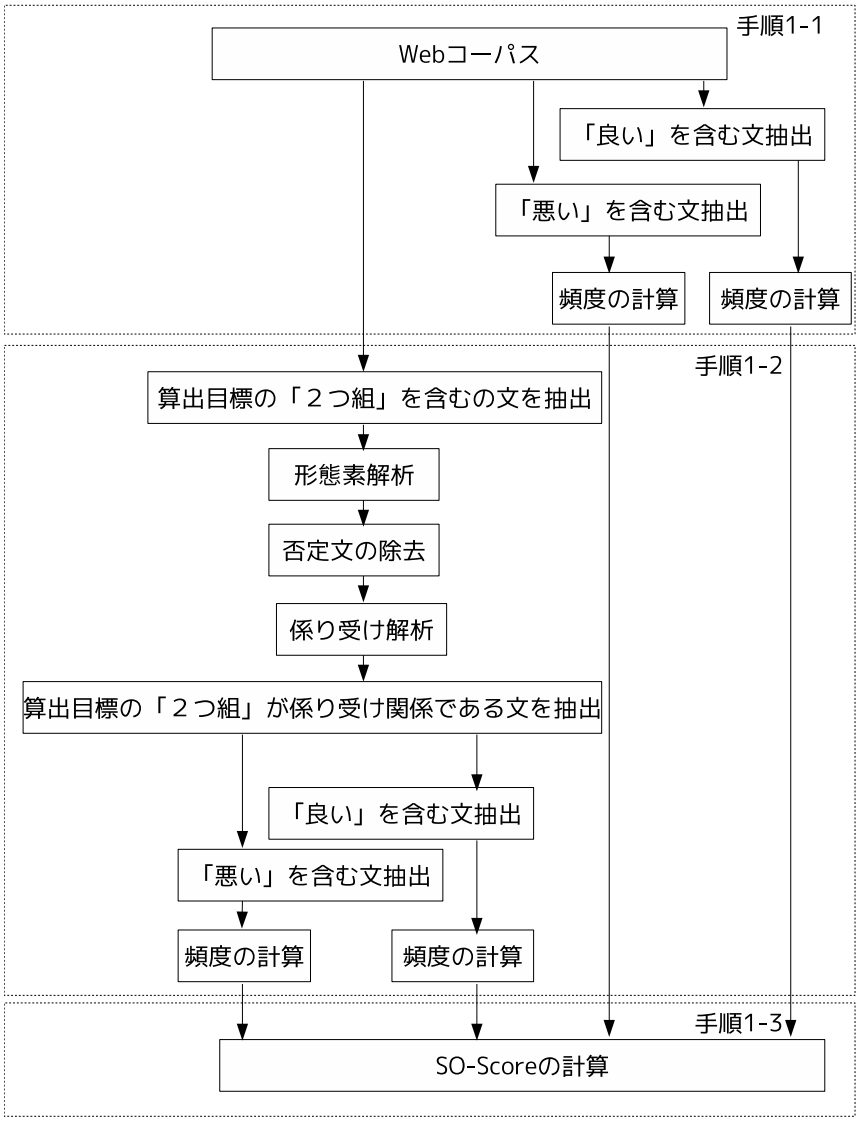

図 4.1: 収集器の実装

4.3 収集結果

手順 2-1 の結果，Positive である文は 221, 527 文，Negative である文は 63, 193 文となった．

手順 2-2 と 2-3 の結果の例を表 4.1，表 4.2，表 4.3，表 4.4 に示す．表 4.1 は Positive の傾向を持つ名詞「写真」について，表 4.2 は Positive の傾向を持つ名詞「店」について，表 4.3 は Negative の傾向を持つ名詞「家」について，表 4.4 は Negative の傾向を持つ名詞「車」についての結果である．表中の順位は，手順 2-2 で抽出した文数による順位である (「良い」,「悪い」と共起しない文を含む)．

着目した 2 つ組については，第 3 章で収集した名詞を使用して，それに対する動詞の 2 つ組を抽出を行った．しかし，5, 748 語の名詞すべてに対する動詞との 2 つ組をコーパスから抽出するにはかなりの実行時間を要してしまうため，名詞ごとの評価極性を *SO-Score* で算出することで，Positive の傾向か，Negative の傾向かの調査を行い，名詞の評価極性の傾向ごとの収集によって考察を行った（名詞ごとの *SO-Score* の算出結果は表 4.5 に示す）．また，抽象度の高い名詞は評価が行うことが困難であるとされているので，具体物となる名詞について収集を行った．ここで「写真」は，*SO-Score* が 1.830 であり，強い Positive の傾向であると考えられる.「店」は，*SO-Score* が 0.911 であり，やや強い Positive の傾向であると考えられる.「家」は，*SO-Score* が -1.014 であり，強い Negative の傾向であると考えられる.「車」は，*SO-Score* が -0.398 であり，やや強い Negative の傾向であると考えられる．

表 4.1: Positive 傾向の名詞「写真」についての 2 つ組の頻度

順位	Pos.	Neg.	*SO-Score*	(写真, 〈動詞〉)
1	640	111	0.718	(写真, 撮る)
2	1,290	364	0.016	(写真, ある)
3	1,344	315	0.283	(写真, する)
4	4,345	856	0.534	(写真, 撮)
5	409	104	0.166	(写真, 見る)
6	868	235	0.075	(写真, あり)
7	158	70	-0.635	(写真, ください)
8	451	112	0.200	(写真, 撮り)
9	132	38	-0.013	(写真, 下さい)
10	559	136	0.230	(写真, なる)
11	47	10	0.422	(写真, 見せる)
12	34	6	0.691	(写真, 入れる)
13	52	24	-0.694	(写真, 写っている)
13	1,078	312	-0.021	(写真, 撮っ)
15	32	8	0.189	(写真, 送る)
15	10	1	1.499	(写真, 取り込む)
15	334	13	2.873	(写真, 撮れる)
15	86	36	-0.554	(写真, 見える)
15	2,226	705	-0.151	(写真, 見)
15	47	21	-0.648	(写真, とる)
21	4	1	0.180	(写真, 加える)
21	43	30	-1.290	(写真, わかる)
23	44	6	1.063	(写真, 作る)
23	28	7	0.189	(写真, 掲載されている)
23	692	96	1.040	(写真, より)
23	263	46	0.705	(写真, できる)
23	7,993	2328	-0.030	(写真, し)
28	36	6	0.773	(写真, 見られる)
28	50	19	-0.414	(写真, 違う)
			- - -	
52	18	2	1.354	(写真, 見れる)
計	38,833	9,707		48,540

表 4.2: Positive 傾向の名詞「店」についての 2 つ組の頻度

順位	Pos.	Neg.	*SO-Score*	(店, 〈動詞〉)
1	3,677	1,068	-0.026	(店, ある)
2	3,836	1,078	0.022	(店, あり)
3	328	126	-0.429	(店, 違う)
4	502	128	0.162	(店, 入る)
5	278	120	-0.598	(店, 出る)
5	36,588	11,262	-0.111	(店, し)
7	166	62	-0.389	(店, 並ぶ)
8	794	178	0.348	(店, 買う)
8	312	34	1.388	(店, 選ぶ)
10	370	58	0.864	(店, 食べる)
11	40	0	10.156	(店, 構える)
11	218	12	2.372	(店, 教えてください)
11	2,044	720	-0.304	(店, なる)
11	2,558	840	-0.203	(店, なり)
11	6,418	2,580	-0.495	(店, いう)
16	10	2	0.507	(店, 連ねる)
16	748	304	-0.511	(店, 入り)
16	290	28	1.562	(店, 探す)
			- - -	
29	234	44	0.601	(店, 作る)
計	90,423	28,294		118,717

表 4.3: Negative 傾向の名詞「家」についての 2 つ組の頻度

順位	Pos.	Neg.	*SO-Score*	(家, 〈動詞〉)
1	5,261	2,353	-0.649	(家, ある)
2	199	31	0.872	(家, 建てる)
3	2,685	1,091	-0.510	(家, あり)
4	249	188	-1.404	(家, 帰る)
5	6,248	3,159	-0.826	(家, いる)
6	264	142	-0.915	(家, 出る)
7	661	162	0.219	(家, 来る)
7	2,181	921	-0.566	(家, なる)
9	515	246	-0.744	(家, 行く)
10	6,302	2,608	-0.537	(家, する)
11	150	51	-0.253	(家, 買う)
12	119	26	0.384	(家, つくる)
13	93	51	-0.943	(家, 住んでいる)
13	112	52	-0.703	(家, 守る)
13	551	148	0.087	(家, 建て)
13	249	141	-0.989	(家, 帰り)
13	25,403	10,696	-0.562	(家, し)
13	353	68	0.566	(家, 作る)
		- - -		
39	335	160	-0.744	(家, なっている)
計	101,942	43,832		145,774

表 4.4: Negative 傾向の名詞「車」についての 2 つ組の頻度

順位	Pos.	Neg.	*SO-Score*	(車, 〈動詞〉)
1	2,929	1,283	-0.619	(車, ある)
2	396	125	-0.146	(車, 止め)
3	433	143	-0.211	(車, 走る)
3	665	234	-0.303	(車, 行く)
3	1,680	699	-0.545	(車, あり)
6	1,180	531	-0.658	(車, なる)
7	196	67	-0.261	(車, 停め)
8	161	51	-0.151	(車, 走らせ)
8	84	38	-0.665	(車, 向かう)
10	1,845	710	-0.432	(車, 走)
11	52	16	-0.110	(車, 走らせる)
11	1,288	635	-0.789	(車, なり)
13	98	29	-0.053	(車, 入れる)
13	488	221	-0.667	(車, 乗る)
13	1,024	582	-0.995	(車, 乗り)
13	150	96	-1.166	(車, 乗せ)
13	69	35	-0.831	(車, つける)
13	3,748	1,597	-0.579	(車, する)
19	35	17	-0.768	(車, 停める)
19	321	96	-0.068	(車, 走り)
19	59	23	-0.451	(車, 止める)
19	107	24	0.346	(車, 行ける)
		- - -		
36	107	81	-1.408	(車, かかる)
計	25,596	11,312		36,908

表 4.5: 名詞の頻度

順位	Pos.	Neg.	*SO-Score*	名詞
1	291	70	-0.280	人
2	165	20	0.708	自分
3	140	0	11.437	日本語
4	93	9	1.0320	手
5	82	19	-0.227	他
6	89	1	4.126	スタッフ
7	76	8	0.911	店
7	73	11	0.394	感じ
9	65	12	0.101	気
10	61	14	-0.213	本
11	59	14	-0.261	目
12	60	11	0.111	内容
13	53	10	0.069	作品
14	54	8	0.418	日本
15	50	10	-0.015	場所
16	46	12	-0.398	車
16	42	16	-0.944	顔
18	54	3	1.830	写真
19	40	16	-1.014	家
20	47	7	0.410	情報
21	47	6	0.632	先生
22	37	14	-0.934	状態
		- - -		
634	5	0	6.633	ＢＢＳ

4.4 評価

4.4.1 規模に関する評価

情緒の直接表現もしくは Positive または Negative の評価極性に関わる文は，第 3 章で因果表現文に着目して 12,060 文を収集したことに対して，本節で共起頻度に基づくことで 284,720 文を収集できた．そのうち，因果表現文によって収集した 1 組の 2 つ組の最大文数が 79 文であることに対し，この手法では，数百倍も多く収集することができている．これにより，2 つ組についての評価極性は，因果表現文では算出できなかったが，共起頻度に基づくことで算出できるようになった．

4.4.2 評価極性の妥当性に対する評価

得られた 2 つ組と評価極性 ($SO\text{-}Score$) への妥当性を考えてみる．表 4.1 によると，(写真, 撮る) は 0.718 であり好評の原因とされる．さらに，(写真, 見れる) は 1.354 であり，より好評の原因とされる．逆に，(写真, 違う) は -0.414 であり不評の原因とされる．(写真, ある) は 0.016 であり比較的中立的といえる．これらは直感的に同意の得られる結果と考える．同様に,「写真」と同じ極性傾向である「店」についても同様な結果が見られる．表 4.2 によると，(店，入る) は 0.162 であり，好評の原因とされる．さらに (店，選ぶ) は 10.156 でより好評の原因とされる．逆に，(店，違う) は -0.429 であり不評の原因とされる．(店，ある) は -0.026 で不評の原因となっているが，(店，あり) では 0.022 となり,「ある」のような動詞は極性を判断することが難しい動詞と考えられる．しかし「ある」「あり」両者はどちらも 0 に近い数値を示していることからも，比較的中立的である結果は直感的に同意の得られる結果と考える．

一方，表 4.3 によると，(家, 建てる) は 0.872 であり好評の原因とされる．これも直感的に同意しやすい結果と考える．しかし，(家, ある) や (家, 帰る) が不評の原因とされている．これは，ブログなどの背景や文脈などを察する必要がある．例えば,「家にある」というとき，忘れ物をしたのかもしれない.「家に帰る」というとき，疲れている文脈かもしれない．Negative 傾向の名詞については分析に注意を要すると考える．同様に,「家」と同じ極性傾向である「車」についても「家」についての結果と同様な傾向が見られる．表 4.4 によると，(車，行ける) は -0.451 であり，好評の原因とされる．これは直感的に同意しやすい結果と考える．しかし，(車，ある) や (車，走る) は不評の原因とされる．例えば,「車にある」というとき，忘れ物をしたのかもしれない.「車が走る」というとき，

燃料が切れそうな状態なのかもしれない．また,「家」に対して,「車」が不評の原因が多いのは「車」が特殊な例であるとも考えられる．また表 4.5 によれば,「車」は「家」と比較するとやや Negative の傾向が低いため，傾向の曖昧性が加算されると考えられる．

第5章　考察

2つ組に対する *SO-Score* の解釈のしやすさと，名詞と動詞の評価極性との関係について考察する．

5.1　名詞と評価極性との関係

名詞の「家」と「車」について2つ組の *SO-Score* を算出したところ，*SO-Score* の解釈が難しかった．これらの名詞はどちらも Negative 傾向であったことから，Negive 傾向の名詞における2つ組は，解釈が難しいと推察される．

そこで，Positive 傾向の名詞 (「写真」と「店」) に対する動詞2つ組と Negative 傾向の名詞 (「家」と「車」) に対する動詞2つ組について，人手による同意率の分析を行った．Positive 傾向の名詞の2つ組は 106 件中 72 件が同意であると判断した．また，Negative 傾向の名詞の2つ組は 99 件中 34 件が同意であると判断した．Positive 傾向の名詞と Negative 傾向の名詞によって，2つ組の評価極性値の同意率に明らかな差があることが判断できる．

解釈の難しさは，パターンに基づく情緒推定 [10] においても同様の現象が見られるので，名詞の評価極性で区別することは重要と思われる．

5.2　動詞と評価極性との関係

動詞について名詞と同様に *SO-Score* を算出してみた．結果を表 5.1 に示す．

表 5.1: 動詞の頻度

順位	Pos.	Neg.	*SO-Score*	名詞
1	1,385	428	-0.642	ある
2	615	10	3.605	できる
3	271	176	-1.713	なる
4	303	107	-0.834	いる
5	194	76	-0.984	あった
6	209	58	-0.487	思っていた
7	169	86	-1.361	する
8	243	1	5.575	出来る
9	140	80	-1.528	違う
10	153	30	0.014	思う
11	148	30	-0.034	出る
12	23	36	-0.563	いう
13	118	27	-0.208	なっている
14	119	23	0.035	なった
15	118	22	0.087	あります
16	116	14	0.714	見える
17	123	3	3.017	できた
18	119	3	2.969	わかる
			- - -	
67	31	10	-0.704	走る
			- - -	
600	5	0	6.633	お送りします

表5.1において,「違う」に対する *SO-Score* は −1.528 である．表4.1，表4.2において，動詞が「違う」については *SO-Score* は負値であった．同様に,「走る」に対する *SO-Score* は，−0.704 である．表4.4において，動詞が「走る」に対する *SO-Score* は負値である．これらのことから，動詞の評価極性によっても2つ組の評価極性値に影響を与えると考えられる．

以上より，今後は，名詞と動詞の評価極性の組み合わせ（4通り）に区別しながらデータ収集を行う必要があると考える．

5.3 残された課題

共起頻度に基づく収集は，網羅性が高いが，Negative 傾向の名詞に対する評価極性の信頼性確保が課題として残った．これには多くの2つ組の収集を行いながら対策を考える必要がある．一方，その点，因果表現文からの収集には，因果という文法的な後ろ盾がある．未使用の接続表現が多く残されているので，その使用により規模拡大の余地がある．ゆえに，2種類の収集方法を効果的に用いることを今後も試みる必要がある．

第6章　おわりに

情緒推定では，原因文の収集自体が重要な課題である．本研究では，Web コーパスから，因果表現文からの収集と共起頻度に基づいた収集の 2 つの方式で情緒生起原因を収集することができた．5 億文から，因果表現文の収集は，Positive の文 10,066 文，Negative の文 1,994 文，合計 12,060 文を収集し，その中から 10,333 の「2 つ組」を得ることができた．共起頻度に基づいた収集では，5 億文から Positive の文 221,527 文，Negative の文 63,193 文，合計 284,720 文を収集し，その中から 205 の「2 つ組」を得ることができた (4 つの名詞「写真」,「店」,「家」，および,「車」についてそれぞれに対する動詞の組合せについて収集)．因果表現文によって収集した「2 つ組」の中は，9 割以上が頻度 1 である組合せであったため，信頼性が確保できなかったが，因果関係の文法構造に着目した収集は必要であると考えられる．また，共起頻度に基づいた収集で得た 2 つ組の評価極性値によって，Negative 傾向の名詞よりも Positive 傾向の名詞については直感的に同意しやすい見解が得られた．

収集方法自体に新規性はないが，情緒推定技術の重要な言語資源の確保，および，言語の意味理解に関する知見獲得のために，今後も収集と考察を続ける必要があると考えている．

謝辞

本研究を進めるに当たり，種々の助言を頂きました村田真樹教授に心から御礼申し上げます．御多忙の中，助言をいただきました松村幸輝教授に心から御礼申し上げます．3年間に渡って御指導いただきました村上仁一准教授に心から御礼申し上げます．徳久雅人講師には，終始に渡り研究の進め方や本論文の書き方など，細部に渡る御指導を頂きました．ここに深く感謝致します．

情緒生起原因の収集のためのリソースとして使用した「5億文 Web コーパス」を提供して下さった河原大輔氏に深く感謝致します．原因文抽出の手がかりとして使用した「評価値表現辞書」を提供して下さった小林のぞみ氏に深く感謝します．本研究を進めるきっかけになった徳久良子氏に敬意を表します．

参考にさせて頂いた文献の著者の方々に対して深く感謝します．本研究にご協力頂いた計算機工学 C 講座の皆様に深く感謝致します．

参考文献

[1] 田中努, 徳久雅人, 村上仁一, 池原悟: “結合価パターンへの情緒生起情報の付与”, 言語処理学会第 10 回年次大会発表論文集, pp.345-348, 2004.

[2] 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.

[3] 徳久雅人, 岡田直之: “パターン理解的手法に基づく知能エージェントの情緒生起”, 情報処理学会論文誌, Vol.39, No.8, pp.2440-2451, 1998.

[4] 滝川晃司, 徳久雅人, 村上仁一, 池原悟: “情緒推定用パターン辞書における荒いレベルの情緒原因判断条件”, 電子情報通信学会技術研究報告, 言語理解とコミュニケーション, NLC2009-40, pp.43-48, 2009.

[5] Daisuke Kawahara and Sadao Kurohashi: “Case Frame Complication from the Web using High-Performance Computing”, In Proceedings of Language Resources and Evaluation, pp.1344-1347, 2006.

[6] 寺村秀夫: “日本語のシンタクスと意味”，くろしお出版，1982

[7] 小林のぞみ, 乾健太郎, 松本裕治, 立石健二, 福島俊一: “意見抽出のための評価表現の収集”, 自然言語処理, Vol.12, No.2, pp.203-222, 2005.

[8] Peter D. Turney: “Thumbs up or thumbs down? Senmantic orientation applied to unsupervised classification of reviews”, In Proceedings of the Association for Computational Linguistics, pp.417, 2002.

[9] 池原悟, 宮崎正弘, 白井諭, 横尾昭男, 中岩浩巳, 小倉健太郎, 大山芳史, 林良彦: “日本語語彙大系”, 岩波書店, 1997.

[10] 野口和樹, 滝川晃司, 徳久雅人: “情緒属性付き結合価パターン辞書により各要素の評価極性を考慮した情緒推定”, 電子情報通信学会技術研究報告, 思考と言語, Vol.111, No.227, TL2011-36, pp.63-68, 2011.