

Web テキストにおける評価情報の整理・分類および タグ付きコーパスの構築

川田 拓也* 中川 哲治† 森井 律子† 宮森 恒†
赤峯 享† 乾 健太郎† 黒橋 禎夫*† 木俣 豊†
* 京都大学, † 情報通信研究機構

1 はじめに

Web 上では誰もがあらゆる話題について情報を発信することができる。それに伴い Web テキストから人々の評価を抽出する手法についての議論が活発になっている。Web 上で評価される対象は、商品やサービスといった利用物や具体物をはじめ、社会や自然に関するあらゆる事象が含まれる。例えば、「少子化」といった社会現象や「地球温暖化」のような環境問題、さらに「消費税増税」のような政策も人々の評価の対象となる。また、最近では、評判分析のように利用物や具体物を対象にしたものだけでなく MPQA コーパス [8] (以下 MPQA) を始め、対象を限定しない評価情報抽出タスクの設計も増えてきている。

我々は『情報分析システム WISDOM』 [1] における評価情報分析エンジンの開発を目的に、与えられたトピックに関連するデータから評価情報を抽出するタスク設計を行ってきた。本稿ではまず、主観・感情的記述に限らず、ある種の客観的記述も評価情報としてみなし、抽出対象とすることを述べる。また、トピックに応じて評価の視点が変わることを指摘し、「評価」には利用物に対する評価、出来事に対する評価、および行為に対する評価があることに留意してタグ付けを行うことを提案する。最後に現仕様に基づき作成したタグ付きコーパスについて簡単に報告する。

2 関連研究

これまでの評価情報タグが付与された大規模なコーパスとして、MPQA [8] や NTCIR-6 における意見分析タスクのコーパス [7] (以下 NTCIR-6) がなどが公開されている。MPQA は新聞記事を対象にしており、文中の ‘Objective speech event’ と呼ばれる事実的言明と ‘private states’ と呼ばれる感情、主観的言明、および主観表現 (expressive subjective elements) についてタグ付与がなされている。それぞれのタグ属性として言明の行為者 (source) と、主観表現であれば態度極性 (positive/negative) などが与えられる仕様である。引用表現によって入れ子になった文では source がそれぞれの文に与えられるため、誰による評価なのか明示される。

NTCIR-6 のタグ付与の仕様は当該文の意見性の有無と意見保持者 (Opinion holder)、与えられたトピックとの関連性と意見極性 (中立も含めて 3 値) が作業者によって付与される。

これまでの問題設計の課題として次の 2 点が挙げられる。

- **客観的記述に含意された評価:** 評価の有無が主観 (意見) 性と客観 (事実) 性の対立で区別されており、客観的な記述から評価が読み取れるケースについて扱う必要がある
- **評価の意味:** 多様な評価対象を扱うために評価対象の意味クラスに応じた評価の意味の差をコーパスに反映させる必要がある

次節では主としてこの二点について議論し、それをふまえて仕様の設計について述べる。

3 「評価」の位置づけ

まず本稿における「評価」の基本的な考え方について述べ、課題となっている「客観的記述に含意された評価」と「評価の意味」について議論する。

3.1 評価要素

評価情報とは、素朴に考えれば、ある対象についての肯定、もしくは否定といった極性を持つ情報を指すと考えられる [2]。また、これまでの先行研究 [3, 6] でも触れられているとおり、評価情報はいくつかの要素から成り立っていると考えられる。我々の仕様では、評価情報を、言語表現として現れている**評価表現**とその**評価保持者**、**評価対象**、および**極性**によって構成された複合的な情報であると想定する。それぞれの要素を**評価要素**と呼び、(1) ように表記する。

- (1) (評価保持者, 評価対象, 評価表現, 極性)

例えば (2a) に対して、(2b) のようにタグ付けを行う。なお、詳細なタグ付け仕様は 4 節で説明する。

- (2) a. 山田博士によるとアガリクスは体に良い。[†]
b. (山田博士, アガリクス, 体に良い, +)

3.2 客観的記述に含意された評価

先述したとおり、これまでの仕様では、評価は主観情報としてみなされてきた。しかし、次に示すように客観的記述の中にも評価情報と捉えられるものはある²。例えば次のような例が挙げられる。

- (3) a. 厚労省は健康食品 A に発ガンを促進する作用が認められたと発表した。[‡]
b. (厚労省, 健康食品 A, 発ガンを促進する作用が認められた, -)
- (4) a. マイナスイオンは病気に対する抵抗力を高めます。[‡]

¹無記号の例文は Web テキストからそのまま引用した実例であることを表す。† が付された例文は作例, ‡ は実例を改変した例を表す。

²この問題点については Nigam ら [5], 大塚ら [6] において ‘evaluative factual’ として紹介されている。

*tkawada@ling.bun.kyoto-u.ac.jp, kuro@i.kyoto-u.ac.jp

†{tnaka, morii, miya, akamine, inui, kidawara}@nict.go.jp

- b. ([著者], マイナスイオン, 病気に対する抵抗力を高めず, +)

(3a) は健康食品 A に対する否定的な評価表現, (4a) はマイナスイオンに対する肯定的な評価表現が含まれていると解釈できる。

このような客観的記述を評価情報とみなすことで「商品 A はダメだ」という抽象的な情報だけではなく、「商品 A は三日で壊れた」、「商品 A の電池は 1 時間で切れた」のような、評価の根拠に関わる情報も捉えることができる。

3.3 評価の意味

Martin [4] によると評価には次の 3 つの側面があると言われている。まず、「好き/嫌い」や喜怒哀楽、欲求など、評価保持者の感情 (AFFECT), 「賢さ」「誠実さ」など物事の善悪や行為についての評価保持者の判断 (JUDGEMENT), もう一つは「美しさ」など評価保持者の感情的、美的な判断 (APPRECIATION) が想定されている。

我々は、APPRECIATION は人工物や事態に対する評価で、JUDGEMENT は行為に対する評価であるという Martin の分類に注目した。Martin の「評価」には客観的表現は含まれていないが、利用物に限らず様々な評価対象を扱う上で、評価対象の性質によって評価の意味が変わることに留意することは重要であると考えられる。

以上をふまえ、我々は評価の意味の違いに応じた評価対象クラスとして次の 3 つを想定した。

1. 利用物

☞ アガリクス, バイオエタノール, マイナスイオン, 保険など

2. 制度・行為

☞ 年金制度, 捕鯨, 首都の移転など

3. 現象・出来事 (社会や自然に関する現象)

☞ 子どもの体力低下, CO2 の増加, 円高など

利用物に対する評価とは、評価保持者がその利用物に対する利便性、有用性および、嗜好などに関して肯定/否定的な感情や判断、経験を述べるものを指す。利用物はサービスなど無形であっても、人間が利用するものであれば「利用物」としてみなす。例えば、(5) は「マイナスイオン」が有用であると主張されているので、利用物に対する評価とみなすことができる。

- (5) a. マイナスイオンは血液をきれいにします。

b. ([著者], マイナスイオン, 血液をきれいにします, +)

制度・行為は人間や社会の意志的な取り決め、行いとしてみなすことができる。制度・行為に対する評価は評価保持者の当該制度・行為に対する賛否が典型的な評価として該当する。例えば (6) のような例が考えられる。例えば、(6) は「捕鯨」という行為について賛成している例である。

- (6) a. 捕鯨を行うことに賛成です。†

b. ([著者], 捕鯨を行うこと, 賛成です, +)

現象・出来事は、制度・行為以外の事態を指す。現象・出来事に対する評価は、評価保持者が当該現象・出来事についての望ましさやそれが期待される現象であるのか否かについて判断が注目される。例えば次のような例が

挙げられる。(7) は著者が「子どもの体力低下」を望ましくない現象とみなしているので、「子どもの体力低下」という現象に対する評価である。

- (7) a. 子どもの体力低下は深刻な問題だ。†

b. ([著者], 子どもの体力低下, 深刻な問題だ, -)

4 タグ付与の指針

本節では実際の対象データへのタグ付与に関する仕様の概略を説明する。

4.1 全般的な指針

評価要素抽出の全般的な指針は次の通りである。

1. タグ付与の対象データは一文単位である。
2. 原則として該当箇所を対象データから直接抜き出す。なお、デス/マス, 終助詞 (ネ/ヨ/ナ/ゾ etc.) および感嘆符 (!/?) も評価要素として取り出す。
3. 該当すると思われる表現が対象データ中に明記されていないが、他の対象データに存在する場合はその部分を抽出し、角カッコ ([]) で囲む。
4. 評価要素が指示代名詞の場合、その先行詞が他の対象データから参照できる場合は指示代名詞と角カッコで囲んだ先行詞を併記する。(8)

- (8) a. (山田氏は評論家として有名だ。) 彼は捕鯨に賛成である。†

b. (彼 [山田], 捕鯨, 賛成だ, +)

4.2 評価保持者

評価保持者については次のような指針に基づいている。

1. 評価情報保持者には個人や組織の他、一般名詞も含まれる。
 2. 対象データの著者と保持者が同一の場合は「[著者]」とする。
 3. 著者が他人の評価を引用しているときは、引用元を評価保持者とする。(9)
 4. 評価保持者が不明もしくは不特定多数の場合は「[不定]」とする。(10)
 5. 著者が他人の評価に対して評価をしている場合、同一文中に評価情報が二つあると考える。(11)
- (9) 山田氏は捕鯨に賛成だと自身のブログで書いていました。†
(保持者 = 「山田」)
- (10) マイナスイオンは体に良いと言われている。†
(保持者 = 「[不定]」)
- (11) a. 山田氏が捕鯨に賛成だという意見を私は許せない。†
b. (山田, 捕鯨, 賛成だ, +)
c. (私 [著者], 山田氏が捕鯨に賛成だという意見, 許せない, -)

4.3 評価対象

評価対象については次のような指針を設けた。

1. 評価対象クラスに留意する。詳しくは後述する。
2. 利用物に対する評価の場合、できるだけ属性も含めて評価対象とする。(12)

表 1: 20 トピック全体の統計

タグ付け文数	4000 文
「関連あり」判定された文	3076 文
「評価あり」判定された文	882 文
「関連あり」かつ「評価あり」の文	820 文
肯定の評価情報数	472 個
否定の評価情報数	427 個

表 2: 関連性・評価有無の一致率

	バイオエタノール	年金制度
関連有り判定:Recall	0.97(150/154)	0.99(163/164)
関連有り判定:Precision	0.94(150/159)	0.96(163/169)
関連有り判定:F-measure	0.96	0.98
評価有り判定:Recall	0.77(50/65)	0.71(36/51)
評価有り判定:Precision	0.75(50/67)	0.86(36/42)
評価有り判定:F-measure	0.76	0.77

3. 比較対象が明記されているときはそれを '<>' で囲み, 評価対象と併記する.(13)
- (12) アガリクスの香りが好きです.†
(評価対象 = 'アガリクスの香り')
- (13) X は Y より優れている.†
(評価対象 = 'X<Y>')

以下では評価対象クラスの認定について補足する. 評価対象は 3.3 で述べた評価対象のクラスに応じた評価をつけることが期待される. ただし, 与えられたトピック自体が評価対象に該当する場合もあるが, トピックに関連した内容が評価対象となることもある. また, トピックに関連した評価対象は, トピックの評価対象クラスと一致するとは限らない. 例えば, 現象・出来事である「CO2 の増加」がトピックとして与えられていた場合, 下のようない「CO2 増加への対策」は賛否が可能な行為を対象とする評価情報とみなすことができる.

- (14) a. 1 日につけっぱなしにしているテレビなどの電力を少なくすることによって, 二酸化炭素の削減に効果がある.
- b. ([著者], 電力を少なくする, 二酸化炭素の削減に効果がある, +)

解釈によっては, (15b), (15c) のように, 行為に対する評価と利用物に対する評価の両方をとれる場合がある. その場合は両方評価情報としてみなして構わない.

- (15) a. 商品 A のデザインを改善することが必要だ.†
- b. ([著者], 商品 A のデザインを改善すること, 必要だ, +)
- c. ([著者], 商品 A のデザイン, 改善することが必要だ, -)

4.4 評価表現

評価表現については下記の要領で行っている.

1. 行為を対象とする評価をとる場合はモダリティ情報にも考慮してタグ付けを行う. 例えば「[P] べきだ」のような場合, 命題部 P を行為に対する評価対象とみなし, 「べきだ」を評価表現とする. (16)
 2. 評価情報が複数ある場合は, 評価対象と極性が同じなら一つだけ取ればよい. (17)
- (16) a. 捕鯨を行うべきではない.†
 - b. ([著者], 捕鯨を行う, べきではない, -)

表 3: 各評価要素の一致率 (主辞の一致による判定)

	バイオエタノール	年金制度
評価表現:Recall	0.64(44/69)	0.54(32/59)
評価表現:Precision	0.64(44/69)	0.76(32/42)
評価表現:F-measure	0.64	0.63
評価保持者一致率	0.68(30/44)	0.94(30/32)
評価対象一致率	0.57(25/44)	0.66(21/32)
評価極性一致率	0.91(40/44)	0.91(29/32)

表 4: 各評価要素の一致率 (完全一致による判定)

	バイオエタノール	年金制度
評価表現:Recall	0.36(25/69)	0.41(24/59)
評価表現:Precision	0.36(25/69)	0.57(24/42)
評価表現:F-measure	0.36	0.48
評価保持者一致率	0.72(18/25)	0.96(23/24)
評価対象一致率	0.60(15/25)	0.75(18/24)
評価極性一致率	0.92(23/25)	0.96(23/24)

- (17) マイナスイオンによって内分泌系や免疫系が活性化してくると病気に対する抵抗力が増加して様々な症状に対してもいい影響を与えてくれます.
(評価情報 = '様々な症状に対してもいい影響を与えてくれます')

4.5 極性

ここで評価情報は原則として肯定もしくは否定的情報に限っているため, 肯定評価であれば '+' を, 否定評価であれば '-' を付与する.

5 コーパスの構築

上記の仕様にに基づき, 実際に評価情報タグ付きコーパスの構築を行った. 現時点で, 20 のトピック (利用物系 10 トピック, 制度系 10 トピック) について各 200 文ずつ, 計 4,000 文からなるコーパスを作成した.

1. 検索エンジン (google) で対象のトピックをキーワードとして検索し, 関連する Web ページを収集する.
2. トピックの単語を含む文とその直後の文の対を, 検索結果の上位ページから 20 個おきに 100 個, 計 200 文を抽出する.
3. トピックの単語を含む文とその直後の文を, 検索結果の上位のページから順番に計 200 文抽出する.
4. 作業者が, タグ付け対象とする文についてトピックとの関連性の有無, 評価有無, 評価要素のタグ付与をする. その際に文脈情報を参照できるように対象データの前後の文も提示する.

作業者は 1 名で, データの一致率を調査するために別の 1 名が 2 トピック (バイオエタノール, 年金制度) についてタグ付けを行った.

表 1 において全文中から抽出された評価情報の数などを示す. なお, 一文中から複数の評価情報を抽出してもよいので, 肯定評価情報数と否定評価情報数の合計と「評価あり」と判定された文の数は一致しない.

表 2~4 において 2 名の作業員間の一致率を調べた結果を示す. これらの表では, 片方の作業員の結果を正解, もう片方の作業員の結果をシステム出力として考えた場合に, 次のようにして計算される Recall, Precision, F-measure が示されている.

- Recall = システムが正しく出力した数 / 正解の数
- Precision = システムが正しく出力した数 / システムが出力した数

- $F\text{-measure} = 2 \times \text{Recall} \times \text{Precision} / (\text{Recall} + \text{Precision})$

一致率に関しては、二人の作業者が共通して抽出した評価表現について、評価保持者、評価対象、および評価極性が一致した割合を示す。

コーパスのタグ付けに際して修飾要素の扱いなど評価表現の抽出範囲に関しては仕様で定めておらず、作業者の判断に任されているため、主辞の一致率に注目している。そのため、文字列が完全に一致したかどうかだけでなく、作業者が抽出した範囲の主辞が一致したかどうかについても評価を行った。

評価の有無に関しては二人の作業者間で、二つのトピックともに8割以上一致している。各評価要素の一致率については主辞の一致率でも極性を除けば概ね5~6割程度にとどまった。

6 課題

本節では、作業者間で一致しなかった事例を挙げながら、現仕様の課題について述べる。

1. 評価を含意する客観的記述の判定 評価が含意される客観的記述の判定は作業者間の解釈にゆだねられ、評価有無の判定の不一致が見られた。下の例では一方の作業者は「バイオ燃料は良いものだ」と保持者が判断したから開発され生産が開始される」と解釈して(18b)としたが、もう一方の作業者は「評価無し」と判定した。

- (18) a. トヨタが2009年からバイオ燃料を開発、生産開始した。[‡]
 b. (トヨタ, バイオ燃料, 開発, 生産開始, +)

2. 評価情報が複数ある場合 下の例のように複数の評価情報が併置されていると考えられ、一方のデータについてしかとれなかった場合が見られた。

- (19) a. 米国はブラジルからのバイオエタノール迂回輸出を問題視しており、(中略) 関税の賦課を正当化している。[‡]
 b. (米国, 関税の賦課, 正当化している, +)
 c. (米国, バイオエタノール迂回輸出, 問題視している, -)

3. 評価対象クラスの認定不一致 次の例は、評価対象クラスの判定が一致しなかった例である。利用物として海藻を評価対象としてみなすか、海藻から生産する行為を評価対象としてみなすかで判定が一致しなかった。仕様では両方とっても良いが、作業者が互いに一方を評価対象として認定したので評価情報が一致しない結果となった。

- (20) a. (中略) 海藻からバイオエタノールを生産すれば将来の国内需要を十分まかなえる。[‡]
 b. ([著者], 海藻, 将来…まかなえる, +)
 c. ([著者], 海藻から…生産す, 将来…まかなえる, +)

4. 評価保持者に関する問題 評価保持者の同定について一致しないケースとして(21)のような例があった。一方の例では国民が不信感を抱いていて著者はそれを報告しているに過ぎないという解釈の結果(21b)のようにタグ付けが行われ、もう一方では、著者自身が制度批判しているという解釈の結果(21c)のようにタグ付けがされたと思われる。

- (21) a. 近年、年金制度に対して国民は強い不信感を抱いている。[‡]
 b. (国民, 年金制度, 強い不信感を抱いている, -)
 c. ([著者], 年金制度, 強い不信感を抱いている, -)

5. 極性のない評価 その他に極性のない評価の扱いが問題として残っている。我々の現在の仕様では極性のないものは評価情報としてみなしていない。しかし極性が決められないにもかかわらず何らかの評価がされているように見える情報もある³。例えば、対象に対する知識が不足していて評価するに至らない場合や、態度を保留する場合などが考えられる。

- (22) a. アガリクスが効くかどうか分かりません。[†]
 b. 捕鯨については今後の成り行きを見ていきたいと思います。[†]

また、トピックに対する興味がない場合や、トピックの前提を否定するような場合も現在の仕様では評価をつけられない。

- (23) a. 地球温暖化には興味ありません。[†]
 b. 地球温暖化なんて嘘です。[†]

このような情報についてどのようにタグ付けを行っていくかは今後の課題となる。

7 おわりに

本稿では、評価情報コーパスを構築するに当たって客観的表現に含意される評価情報に注目すること、および評価対象クラスに応じた評価の性質に留意して評価情報タグの付与を行うことを提案した。今後は残された課題の解決および作業者間の一致率を上げるため、より精緻な仕様の設計を行っていく予定である。

参考文献

- [1] 赤峯亨, 宮森恒, 加藤義清, 中川哲治, 乾健太郎, 黒橋禎夫, 木俣豊: Web情報の信頼性検証のための情報分析システム WISDOM, 言語処理学会第14回年次大会論文集 [本論文集] (2008).
 [2] Hunston, S.: Counting the uncountable: Problems of identifying evaluation in a text and in a corpus, *Corpora and Discourse* (Partington, A., Morley, J. and Haarman, L.(eds.)), Peter Lang, Bern, pp. 157-189 (2004).
 [3] Kobayashi, N., Inui, K. and Matsumoto, Y.: Opinion Mining from Web Documents: Extraction and Structurization, 人工知能学会論文誌, Vol. 22, No. 2, pp. 227-238 (2007).
 [4] Martin, J. R. and White, P. R. R.: *The Language of Evaluation. Appraisal in English*, Palgrave Macmillan, Houndmills (2005).
 [5] Nigam, K. and Hurst, M.: Towards a Robust Metric of Opinion, *AAAI Spring Symposium on Exploring Attitude and Affect in Text*, pp. 27-34 (2004).
 [6] 大塚祐子, 乾孝司, 奥村学: 意見分析エンジン -計算言語学と社会学の接点-, コロナ社, 東京 (2007).
 [7] 関洋平, Evans, D. K., Chen, H.-H., Ku, L.-W., 神門典子: 意見分析タスク -多言語テキストを対象とした意見抽出技術の評価-, 情報処理学会 研究報告, 2008-FI-89 (8), 2008-NL-183(8), pp. 51-57 (2008).
 [8] Wiebe, J., Wilson, T. and Cardie, C.: Annotating expressions of opinions and emotions in language., *Language Resources and Evaluation*, Vol. 39, No. 2-3, pp. 165-210 (2005).

³ 大塚ら [6] における「評価を含むが極性のない中立的クラス」にあたる。