

タグ付けツールによる評判情報コーパスの作成支援

唯野良介† 嶋田和孝† 遠藤 勉†

九州工業大学情報工学部知能情報工学科†

1 はじめに

近年、Weblog や Web 掲示板等の普及により、個人が製品やサービスなどに対するレビューを投稿する機会が増えている。これに伴い、自然言語処理において、レビュー等の評価情報を対象とした分析(評判分析)の重要性が増している。この評判分析の研究における重要なタスクとして、評価情報コーパスの構築が挙げられる。評価情報のモデルは様々であるが、例えば小林ら [1] は評価情報を(根拠、評価、態度)という意見の構成要素に分類してタグ付けを行い、意見タグ付きコーパスを作成している。本研究においても、評価表現とその極性(肯定/否定)のペアに加えて、評価項目という情報を考慮したタグ付きコーパスの作成を行っている。しかし、人手によるコーパスの作成には多大な手間と時間がかかり、またコーパスの質といった問題も問われる。

そこで本稿では、質の高いコーパスを効率的に作成するためのシステムを構築し、タグ付きコーパスの作成支援を行う。

2 使用データ

本研究での対象は、Shimada ら [2] が評価文書分類で使用したゲームレビュー文書である。この評価文書には、各レビュー記事ごとに7つの評価項目が含まれている。評価項目は「オリジナリティ(o)」、「グラフィック(g)」、「音楽(m)」、「熱中度(a)」、「満足度(s)」、「快適度(c)」、「難易度(d)」である。この文書の一部(約5,000文)に対して2名のアノテータ(A_1, A_2)がそれぞれタグ付けを行った。アノテータは文書内から評価表現を選別し、タグを付与する。タグの種類は極性タグと、選んだ評価表現がどの評価項目に関連するかのタグの計2種類である。タグ付けした文章数はそれぞれ A_1 :3,446 文、 A_2 :1,589 文である。互いの文章数に差があるのは、 A_1 はこのデータを他研究で使用した際に何度か修正・追加を行ったためである。実際のタグ付け例を図1に示す。

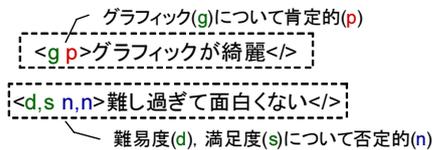


図1 タグ付け例

2人のアノテータのタグ付け結果が一致しているほど、作成したコーパスは信頼できるといえる。そこで、 A_1 と A_2 のタグ付けの一致度を算出したところ、両ア

ノテータが同一箇所の表現にタグを付与した割合(評価表現の選択一致率)は0.427、付与したタグが互いに一致した割合は κ 値で0.456であった。この値は κ 値の指標において中程度の一致度といえるが、十分な値とはいえない。また、アノテータによる評価表現の見落としも多数存在していることがわかった。

3 タグ付け支援ツール

本節では2節で得られた結果を踏まえて、より効率的にタグ付きコーパスを作成するための支援ツールを構築する。ツールに現在実装されている機能は次の3つである。図2に実際のツール画面を示す。



図2 支援ツール画面

3.1 GUIによるタグ付け操作の簡易化

2節では既存のエディタを用いてコーパスを作成した。しかし、タグ付け作業に手間がかかり、アノテータにかかる負荷の増大、コーパスの作成効率低下といった問題が生じてしまった。そこで、タグ付け操作を簡易化するためのGUIを作成する。これにより、マウス&クリック等の簡易作業でタグ付けを行うことが可能になる。

3.2 評価項目と関連度の高い表現の強調表示

タグが付与されやすいと思われる表現をあらかじめ強調表示させることで、タグ付けにかかるコストを減少させる。そのためにまず、2節で準備したタグ付きコーパスから各評価項目と関連度の高い表現の抽出を行う。例えば、「グラフィック」の評価項目タグが付与された表現中には「画像」という単語が頻出するという傾向がある。この場合、「グラフィック」と「画像」は関連度が高いとし、この2つをペアにして抽出する。今回はtf-idfを用いることで評価項目別に関連度の高い単語を抽出する。また単語の頻度情報を用いて、極性と単語のペアも抽出する。このようにして得られた表現を用いて、文書内の同一表現を強調表示(文字色の変更)する。

An annotation support tool for building a sentiment corpus
†Ryosuke Tadano, Kazutaka Shimada, and Tsutomu Endo
†Department of Artificial Intelligence, Kyushu Institute of Technology

3.3 タグ付き事例の提示

信頼度の高いコーパスを作成するためには、付与したタグがアノテータ間で異なる「タグ揺れ」の問題を解消する必要がある。宮崎ら [3] はそのための手法として、過去のタグ付き事例の提示を行っている。タグ付き事例を参照することでアノテータ間の判断を共有させることが可能となる。そこで本ツールにおいても事例提示を行うことで、タグ揺れ問題の解消を試みる。提示する事例の種類は以下の2つである。

- 別文書からの類似事例の提示
アノテータが対象とする表現に対して、過去のタグ付き事例集の中から類似度の高いものを提示する。
- 同一文書からのタグ付き事例の提示
同一文書に対する他アノテータのタグ付け例を提示する。これはアノテータが各自タグ付けを行った後、見直しをする際に使うものである*1。

4 評価実験

3節で構築した支援ツールを用いて実際にコーパスを作成し、支援ツールの有効性を確かめた。対象データはゲームレビュー文書(2節とは異なるデータ)とし、アノテータは2名(T_1, T_2)である。ツールの各機能の効果をそれぞれ調べるために、同じレビュー文書に対して次の4つのケースで順にタグ付けを行った。

- ケース1. 支援ツール無し
 ケース2. GUI有り + 強調表示あり
 ケース3. ケース2 + 類似事例提示あり
 ケース4. タグ付き事例あり(ケース3の見直し)

ケース3の提示事例として、2節で作成したコーパスの中で A_1 と A_2 のタグ付け結果が完全に一致した事例(1,109文)を用いた。実験はデータを変えて2度行った。順番に実験1, 実験2とし、実験結果を表1から表4に示す。表1, 表3の項目は、「評価表現数」はアノテータがタグ付けした表現の総数, 「付与タグ数」はアノテータが表現に付与したタグの総数, 「作業時間」はタグ付けにかかった時間を表す。なお、実験に用いたレビュー文書は実験1: 約350文, 実験2: 約450文であり、実験2ではケース1を省略している。

表2, 表4の実験結果から、使用する機能が増えるにつれてタグ付けの一致度が上昇していることがわかる。この結果から、今回作成した支援ツールはタグ揺れの減少に有効であるといえる。特に、ケース4における結果が最も良かったことから、同一文書からのタグ付き事例の提示は見直し作業として有効であることがわかる。

ケース3の事例提示による κ 値の上昇率は十分でなかった。これは用意した事例の量が少なく、類似事例として適切な表現を提示できていないことが原因だと考えられる。また、ケース3において実験1, 2共に評価表現の抽出数、選択一致率が下がる結果となった。これより、事例提示による支援は、評価表現の抽出にも影響を及ぼすことがわかった。しかしその場合でも、明らかな評価表現の見落としはほとんどなく、対象データの量と

表1 各アノテータのタグ付け結果(実験1)

ケース		1	2	3	4
T_1	評価表現数	323	330	325	342
	付与タグ数	412	423	414	426
	作業時間(分)	73	43	64	51
T_2	評価表現数	382	384	365	374
	付与タグ数	409	431	433	482
	作業時間(分)	120	95	125	60

表2 アノテータ間のタグ付けの一致度(実験1)

ケース	1	2	3	4
評価表現の選択一致率	0.828	0.824	0.808	0.874
付与したタグの一致度(κ 値)	0.350	0.419	0.442	0.573

表3 各アノテータのタグ付け結果(実験2)

ケース		1	2	3	4
T_1	評価表現数	—	415	395	415
	付与タグ数	—	526	508	526
	作業時間(分)	—	70	99	58
T_2	評価表現数	—	464	462	480
	付与タグ数	—	530	536	598
	作業時間(分)	—	118	148	67

表4 アノテータ間のタグ付けの一致度(実験2)

ケース	1	2	3	4
評価表現の選択一致率	—	0.840	0.815	0.857
付与したタグの一致度(κ 値)	—	0.491	0.526	0.687

比較しても十分な量をタグ付けできているといえる。

全体的に実験1よりも実験2の方がタグ付けの一致度が高い結果となった。これは実験1のケース4において、他のアノテータのタグ付け例を一度参照したことでタグ付けの判断の共有が行われたためだと考えられる。これより、アノテータがこの後互いに異なる文書に対してタグ付けをした場合でも、タグ揺れの少ないコーパスの作成が期待できる。

5 おわりに

本稿ではタグ付けツールを試作し、ツールによる評判情報コーパスの作成支援を行った。コーパスの作成実験を行った結果、アノテータ間のタグ付けの一致度を上げることができ、ツールの有効性を確認できた。今後は新たな支援方法を検討し、より質の高いコーパスを効率的に作成できるシステムの構築を目指す。

参考文献

- [1] 小林のぞみ, 乾健太郎, 松本祐治. 意見情報の抽出/構造化のタスク仕様に関する考察. 情報処理学会研究報告, NL171-18, pp.111-118, 2006.
- [2] K. Shimada and T. Endo. Seeing several stars: a rating inference task for a document containing several evaluation criteria. Proceedings of PAKDD 2008, pp.1006-1014, 2008.
- [3] 宮崎林太郎, 前田直人, 森辰則. 人手による評判情報注釈付けにおける揺れの分析と注釈付け支援ツール. 情報処理学会研究報告, NL176-21, pp.143-150, 2006.

*1 予備実験から、初めからこの機能を使用するとアノテータの視野が狭まり、評価表現を見落とししてしまうことがわかったため。