社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2006－NL－176 (21)
2006／11／23

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

宮崎林太郎
横浜国立大学 大学院
環境情報学府

前田直人
横浜国立大学 大学院
環境情報学府

森辰則
横浜国立大学 大学院
環境情報研究院

E-mail:{ rintaro, n-maeda11, mori}@forest.eis.ynu.ac.jp

本稿では，人手による評判情報コーパスの作成における注釈事例参照の利用の提案と，注釈事例参照を組み込んだ注釈付け支援ツールの試作実験を報告する．

複数注釈者による人手の注釈付けには，注釈揺れが必ず存在する．そこで，注釈事例の参照を行うことで，複数注釈者間の認識の違いの削減する手法を提案した．同時に，注釈付けツールを用いることで注釈作業を簡単にした．

評価実験によれば，注釈付けツールを用いて過去の注釈付け事例を参照しながら注釈付けを行った場合と，ツールを用いなかった場合の一致度を比較したところ，平均 0.08，最大で 0.15 程度 κ 値が向上した．また，注釈揺れの内容について分析を行った．

Analysis of Manual Annotation of Sentiment Information in Text and an Annotation Supporting Tool

Rintaro MIYAZAKI　Naoto MAEDA　Tatsunori MORI

Graduate School of Environment and Information Science,
Yokohama National University

In this paper, we propose a way to improve the quality of manually-annotated corpus for sentiment analysis by consulting example sentences that have been already annotated. We also introduce an annotation tool with which human annotators can easily refer to related examples.

In the case of corporative work by plural annotators, the results by different annotators are not always consistent with each other. Therefore, we introduce a method to decrease the disagreement by using annotated examples. We also developed a tool for supporting annotation.

The experimental result shows that the kappa value is improved (0.08 in average, maximum 0.15), when using the annotation supporting tool compeering with the case that the tool is not used. We also conducted a detailed analysis about the disagreement of annotation among different annotators.

1．はじめに

近年，Weblog 等の普及により，個人が製品やサービス等に対するレビューを公開する機会が増えてきている．また，価格.com[1]や Amazon[2]のような，製品レビューを集約して扱うサイトも広く利用されるようになってきている．

上記のようなネット上に存在するレビューには，製品のスペックや数値情報等の客観的な情報に加え，レビューの筆者による主観的な見解が多く含まれている．これを評判情報として収集・分析することが期待されている．

評判情報に関する研究においては，レビュー中の評判情報記述部分を特定する問題を扱う抽出に関する研究と，筆者の意見が肯定的か否定的かを判断する問題を扱う分類に関する研究の二つが大きく扱われている．

本研究の目的は，評判情報抽出タスクに関する研究を推進するにあたって，必要不可欠と考えられる評判情報コーパスを効率よく，かつ精度良く作成することにある．しかし，既存の評判情報に関する研究において，コーパスの質に言及されているものは少ない．

本稿では，複数注釈者が人手により評判情報コーパスを作成する際の注釈揺れの分析を行う．加えて，揺れを減らすための方法として，過去の注

[1] http://www.kakaku.com
[2] http://www.amazon.co.jp

釈事例を参照しながらの注釈付けを提案する．そのために，注釈事例の参照を行いながら注釈付けが可能なツールを試作した．

2．関連研究

2．1 評判情報の抽出

評判情報の抽出に関する研究には大きく分けて2種類のアプローチが存在する．ひとつは文や文書などのまとまった単位で意見性を判定するものである．もうひとつは評判情報を構成する要素を定義し，各構成要素を抽出しようとするものである．

前者の中には文書分類に関する研究として，Turney [2] がある．この研究では，文書中に含まれる評価表現の出現比率から文書全体の評価の極性を求めている．また，文単位での意見性の判定を行ったものとして，峠ら [5] がある．この研究では，文中に現れる単語が意見文になりやすい単語であるか否かを学習し，Web 掲示板から意見文の抽出を行っている．さらに，評判情報の抽出に関するさきがけの研究でもある立石ら [1] では，あらかじめ用意された評価表現辞書を用いて，対象物と評価表現を含む一定の範囲を，意見として抽出している．

一方，後者の中には村野ら [3] がある．この研究では，評価文の文型パターンを整理し，その構成要素ごとに辞書を用意し抽出を行っている．他にも，小林ら [7] は評判情報を＜対象，属性，属性値＞の3つ組として，ブートストラップ的な手法で各要素の収集を行っている．

さらに，評判情報が構成要素からなるとする研究においては，各構成要素の組を同定するという研究も行われている．飯田ら [4] は機械学習手法を用いて属性-評価値対を同定している．

機械学習手法は高村ら [8] などが肯定否定分類に用いており，我々はこのような機械学習手法における訓練データや，他の評判情報抽出手法における，評価データなどに用いるために評判情報コーパスの必要性が高まっていると考えられる．

2．2 評判情報のモデル化

評判情報のモデル化に関する研究においても，村野ら [3] が挙げられる．この研究は，評判情報の構成要素を最も細かく分類したものの一つである．主観的評価の構成要素を [対象] [比較対象] [評価] [項目] [様態] と分類している．

また最近では小林ら [9] が，意見情報の構成要素を [態度] [評価] [根拠] [理由関係] とする研究を行っている．この研究では [評価] についてはさらに細かく，[評価主体] [評価対象] [評価視点] [比較対象] [評価条件] の構成素からなると考え，コーパスの作成を行っている．また，この研究ではコーパスの質についても言及されている．評価値候補に対する 2 名の注釈者のタグ付与の判断がある程度一致したと報告している．

2．3 注釈付け支援ツールと注釈事例の提示

注釈付け支援ツールに関する研究では，高橋ら [11] がユーザの定義したチャンクに対するラベル付けや関係の付与を目的としたアノテーションツールである “Tagrin” を作成している．

また，注釈付け作業における事例の提示に関しては洪ら [6] がある．この研究では対話コーパスの作成の際に，注釈者へ過去のタグ付け事例を提示することで支援を行っている．

他にも，翻訳の分野においては翻訳メモリと呼ばれる原文と訳文のデータベースを用いた翻訳支援が行われている．これも，ある種の事例参照であると言える．

3．提案手法

本章では提案手法について説明する．

本研究の目的は，人手による評判情報コーパスの作成である．しかし，ある程度の規模を持つコーパスを人手で作成する際には複数注釈者による作業の並列化が必要不可欠である．しかし，複数注釈者による注釈付けを行う場合には注釈者による注釈揺れの問題を避けて通ることはできない．

もちろん，完全に一致をするのは不可能だが，各注釈者が独自の判断で注釈付けを行っていては揺れが大きくなる一方である．そこで，複数の注釈者による注釈付けにおいて，揺れを一定の範囲内に収めるための手法を考える必要がある．

我々は，評判情報の注釈付けにおいてどのような揺れが発生するのかを調べた．その結果から，過去の注釈事例を参照しながら注釈付けを行う手法を提案する．これは、複数注釈者間で判断をゆるやかに共有することを目的とするものである。

先行研究に対して，本稿の貢献は，1) 4つ組の要素とオントロジー情報からなる評判情報モデルの注釈付けを行う，2) 注釈付け支援ツールに関する先行研究を参考にし，あらかじめ用意した辞書などを用いずに，複数注釈者の判断を共有することで注釈揺れの削減を目指す点にある．

本節では，最初に評判情報注釈付けのタグセットの説明，次に予備実験の紹介，最後に注釈事例の参照を行いながら注釈付けを行うためのツールを紹介する．

3．1 評判情報のタグセット

最初に，本稿で用いる評判情報のタグセットについて説明する．我々は評判情報の注釈付けタグセットの提案を行い，構成要素を定義してきた [10]．本稿ではそれに従い，評判情報の基本構成要素を図1に示した各項の組であると定義する．

また，今回用いた評判情報タグセットを図 2 に示す．図 2 には図 1 に示した 4 項目に対応するタグと，XML タグの属性情報が示してある．図 2 中の，<>で囲われているのが基本構成要素 4 項目のタグである．その下に列挙されているのが，各 XML タグの属性である．また、全てのタグに対して，そのタグを一意に参照できる id 属性値を付与する．

[項目]：製品やサービスを構成する要素を意味する概念クラスやそのインスタンスを指示する表層表現．
[属性]：項目の様態を表す観点．
[属性値]：属性に対する様態の内容．
[評価]：[項目]に対する主観的見解．

図 1．本稿で扱う評判情報の基本構成要素

<item>：項目
　class：この項目が属する概念番号
<attribute>：属性
　pair：属性-属性値の組番号
　target：属性-属性値を持つ概念番号
<value>：属性値
　pair：属性-属性値の組番号
　target：属性-属性値を持つ概念番号
<evaluation>：評価
　target：評価の対象となる概念番号
　reason：評価理由となる属性-属性値の組番号
　orientation：評価が肯定・中立・否定のどれか

図 2．注釈付けに用いたタグセット

さらに，item 属する概念クラスは部分-全体関係や上位-下位関係といった階層構造を有する．この情報はいわゆるオントロジー (の一部) を構成するものであるので，必ずしも文中に陽に現れるものではないために，我々はオントロジーに関する情報を文中に付与するタグとは別に記述する手法を提案した[10]．記述されたオントロジー情報の例を図 3 に示す．

[c01]<非電化製品>
+[c141]<圧力鍋>
++[c02]<この圧力鍋>
+++[c931]<付録>
+++[c1201]<蓋><ふた>

図 3．オントロジー情報の実例

オントロジー情報はオントロジーの木構造に表れる各ノードを先順で記述してあり，行頭にある+は深さを，その後ろの[]の中が各階層の概念番号を表している．<>の中は表層表現であり，同じ階層に属する表層表現が列挙してある．

3．2 注釈付け予備実験

我々は最初に，複数注釈者による注釈付けにおける問題点を明らかにするために，予備実験を行った．予備実験では一致の度合いを高くする目的で、3 種類の条件で注釈者に注釈付けを行ってもらった．その結果として，各条件付けだけでは注釈者間の一致度が不十分であることがわかった．そのために，今回提案する過去の注釈事例の参照を行いながら注釈付けをする手法を検討した．

予備実験の内容と結果については 4．1 節と 4．2 節で詳しく述べる．

3．2 注釈付けツール

本稿で我々が提案するものは，評判情報コーパス作成における注釈者への注釈事例の提示である．

既存のエディタや注釈付けツールでは今回の実験で必要な事例の提示は行えないので，新たに注釈付けツールを試作した．図 4 はその注釈付けツールの画面である．

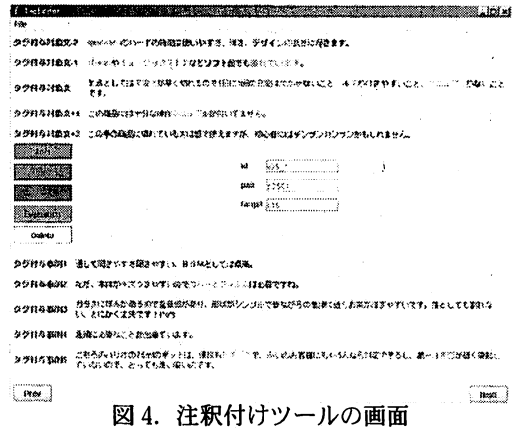

図 4．注釈付けツールの画面

上段部には注釈付け対象となる文とその前後の文を表示している．中段部にはタグ付与のためのボタンと，XML タグの属性値情報入力のためのフォームが用意されている．また，下段部には編集中の文に類似した注釈付け済みの文を例示している．

試作した注釈付けツールの特徴は文単位で注釈事例を提示することである．そのためには注釈付け対象文に対して，事例集の中から類似した注釈事例を探してくる必要がある．次節の実験では，注釈事例の提示を行うために次式を類似度計算に用いた．この類似度は非常に簡単なものであるので、今後検討する余地がある。

$$\text{類似度} = \frac{\text{両文でのBigramの一致数}}{\text{対象文の長さ+事例文の長さ}}$$

現在の設定では、新たな注釈付け対象文が読み込まれる度に，事例集として指定されたファイルに含まれる文の中から注釈付け対象文との類似度が高い，上位 5 文を検索してきて表示する．

また，注釈事例の提示に併せて，注釈付け対象文中の注釈付けされやすいと思われる文字列を

強調表示するようにした．これは注釈付けの見落としを少なくするためである．今回は，注釈付け対象文と提示事例の Bigram が一致した部分を強調表示した．

上記に加え，注釈者の作業を軽減するためボタンによるタグの付与と，各タグの id 属性値の自動入力が行えるように実装した．

4．実験と評価

4．1 予備実験

4．1．1 予備実験の手順

本節では予備実験の内容について述べる．

Amazon から収集したレビュー文書に対して複数注釈者に3．1節で紹介したタグを人手で付与してもらった．

注釈付け対象とした製品のジャンルは，電化製品(Topic1)，非電化製品(Topic2)，映像・音楽(Topic3)，ソフトウェア(Topic4)の4トピックである．各ジャンルについて50文，計200文を実験毎に用意した．

条件を変えて3回実験を行った．実験の内容は次の通りである．

予備実験1：評判の基本構成要素だけでなく，周辺情報も含めて，注釈者に注釈付けを行ってもらった．同時にオントロジー情報も各注釈者に作成してもらう．注釈を行った被験者は情報工学を専攻する学生4名である．2名は本研究との直接の関係を持つが，他の2名は本研究と直接の関係を持たない．

予備実験2：itemタグが既に付与されている文書に対して，評判情報の基本構成要素のみを注釈者に注釈付けしてもらった．注釈を行った被験者は情報工学を専攻する学生5名である．2名は本研究との直接の関係を持ち、予備実験1にも参加している．他の3名は本研究と直接の関係を持たず，予備実験1には参加していない．

予備実験3：item タグが既に付与されている文書に対して，attribute-value の組については片方の要素が省略されている場合には補完をしながら注釈付けを行ってもらった．これにより，注釈者による attribute-value の組の認識が明確になることを期待した．注釈を行った被験者は本研究と直接の関係がある情報工学を専攻する学生2名であり，予備実験1と予備実験2に参加している．

実際には，最初に予備実験1を行い，その結果を見て条件を加えたのが予備実験2である．予備実験3も同様に予備実験2の結果を見て条件を追加した．

4．1．2 実験の結果

予備実験1の結果を表1に示す．attribute(att)，value(val)，evaluation(eval)の3要素について一致率を調べた．尺度にはκ値を用いた．注釈者2名の組（注釈者4名なので全6組）の中で最も高い値と低い値を示してある．

表1．予備実験1の注釈付けの一致率（κ値）

	Topic1	Topic2	Topic3	Topic4
att	0.55	0.69	0.38	0.56
	0.33	0.28	0.11	0.14
val	0.55	0.40	0.37	0.48
	0.43	0.17	0.11	0.16
eval	0.81	0.73	0.42	0.70
	0.51	0.29	0.22	0.37

κ値は主観が入る判定が偶然に拠らず一致する割合であり，0.41から0.6の間ならば中程度の一致，0.8を超えるとほぼ完璧な一致と考えられる．

一部κが0.8を超える高い一致を示している部分もあるが，全体的には低い一致になっている．しかも，これは特定の注釈者が関係した場合にのみ一致率が低くなっているわけではない．どの注釈者の組でも一致率が低くなる可能性があることがわかった．

さらに，揺れの内容を細かく調べた結果，attributeの一致率が低くなっているのは，itemの部分-全体関係とattributeの区別が注釈者によって異なっているためであることがわかった．異なりの例を図5に示す．

注釈者A:<item>逆光補正など</item><value>簡単に行えます</value>
注釈者B：<attribute>逆光補正など</attribute><value>簡単に行えます</value>

図5．item-attributeの揺れの例

図の例では「逆光補正など」という製品の機能を部分と考えるか，属性と考えるかで認識に異なりがあった．

そこで，item の部分全体関係であるオントロジー情報をあらかじめ一人の注釈者が作成し，itemタグを付与した状態で注釈付けを行ったのが予備実験2である．

予備実験2の結果を表2に示す．

表2．予備実験2の注釈付けの一致率（κ値）

	Topic1	Topic2	Topic3	Topic4
att	0.73	0.83	0.5	0.75
	0.46	0.29	0.05	0.40
val	0.93	0.67	0.57	0.64
	0.42	0.19	0.12	0.30
eval	0.65	0.46	0.65	0.59
	0.31	0.23	0.31	0.26

予備実験1と同様に2名の注釈者の組（全10組）のκ値を調べ，その中で最も高い値と最も低い値を示してある．一致度の高い注釈者の組ではκで0.6以上の値となっている部分も多数あり，

ある程度の一致率となっていると考えられる．しかし，一致率の低い注釈者間ではκ値は0.3以下となってしまっている．予備実験1の結果に比べると全体的に高い一致率となっているが，一致率の低い部分に関してはまだ対策を考える必要がある．

次に，我々は注釈揺れを解消する方法として省略されている要素の補完を考えた．レビュー中には評判情報の構成要素が省略されている場合がある．特に，attribute-value 組については片方が省略されている場合が多く存在する．例を図6に示す．

〈value〉小さく〈/value〉て気に入っています．

図6．attributeが省略されている例

図の例文では属性値「小さく」が単独で出現している．これは本来ならばattributeとして現れるべき「大きさ」や「サイズ」といった表層が省略されているものと考えられる．

我々は，注釈揺れの原因の一つにこの省略があるのではないかと考えた．そこで，予備実験3では attribute-value 組の省略を補完しながらの注釈付けを行ってもらった．予備実験3を行った結果を表3に示す．予備実験3は注釈者が2名だったため，2名の間のκ値のみを記してある．

表3．予備実験3の注釈付けの一致率（κ値）

	Topic1	Topic2	Topic3	Topic4
att	0.72	0.44	0.38	0.51
val	0.66	0.48	0.24	0.56
eval	0.77	0.53	0.41	0.49

結果を見ると，多くの場合でκ値が0.4を超えている．このことから，中程度の一致はしていると言える．しかし，予備実験2の最低値と比較すれば一致率は良いが，最高値と比べると一致率が安定しているとはいえず，十分であるとは言えない．この結果から，省略の補完という手法が十分に有用であったとは考えられない．

4．2 注釈付けツールを用いた実験

4．2．1 実験手順

予備実験と同様にAmazonから収集したレビュー文書，4ジャンル200文について注釈付けを行った．予備実験2と同様に製品についての部分全体関係を記述したオントロジー情報があらかじめ作成してあり，item タグがあらかじめ付与されている文書を使用した．今回の実験では，次の三つの条件で注釈付けを行ってもらった．

ツール無し：注釈付けツールを使用しない．注釈事例の提示も行わない．

事例無し：注釈付けツールを使用する．しかし，注釈事例の提示は行わない．

事例有り：注釈付けツールを使用する．注釈事例の提示を行う．

3つの条件の比較を行うため，6人の注釈者を3つのグループに分けて各データについて条件を変えて3回注釈付けを行ってもらった．注釈者のグループ分けと注釈付けの順番を表4に示す．

また，今回事例集として使用したデータは，予備実験の過程で注釈付けされたデータの中から一人分を用いた．

表4．注釈付けツールを用いた実験の手順

	注釈者1･2	注釈者3･4	注釈者5･6
1回目	ツール無し	事例無し	事例有り
2回目	事例無し	ツール無し	事例無し
3回目	事例有り	事例有り	ツール無し

表4のように，注釈者1から注釈者4までには事例有りを最後に行ってもらい，注釈者5･6には事例有りを最初に行ってもらった．

また，今回の実験で注釈を行った被験者は情報工学を専攻する学生6名である．2名（注釈者3と注釈者4）は本研究との直接の関係を持ち、すべての予備実験に参加している．他の4名は本研究と直接の関係を持たず，2名（注釈者1と注釈者2）は予備実験に参加していない．残りの2名のうち1名（注釈者5）は予備実験1に参加しており，もう1名(注釈者6)は予備実験2に参加している．

注釈者には，今回使用した評判情報のモデルについての解説と，注釈事例を用いた注釈付けの説明を行った．

4．2．2 実験結果

本節では，それぞれの条件における実験結果を示す．

最初に，各注釈者が注釈付けしたタグの数を調べた．結果を表5に示す．

表5．各注釈者が付与したタグの数

	ツール無し		事例無し		事例有り	
	1回目		2回目		3回目	
	注釈者		注釈者		注釈者	
	1	2	1	2	1	2
att	86	55	64	74	*111*	*77*
val	127	142	147	153	*199*	*170*
eval	114	*72*	*135*	70	97	*72*
	2回目		1回目		3回目	
	注釈者		注釈者		注釈者	
	3	4	3	4	3	4
att	104	69	89	62	*137*	*82*
val	184	114	181	115	*215*	*126*
eval	53	79	53	81	*81*	*88*
	3回目		2回目		1回目	
	注釈者		注釈者		注釈者	
	5	6	5	6	5	6
att	*86*	66	81	*97*	*86*	83
val	176	204	*211*	*206*	177	181
eval	66	79	54	*85*	*66*	83

多くの注釈者が事例有りの場合に最も多くの注釈付けを行った．しかし，注釈者5と注釈者6については事例無しの場合に最も多くの注釈付けを行っている．このことから，注釈付けの数については注釈者の学習による影響が懸念されるが，注釈者1･2の場合のように，後から行った事例有りの場合の方が注釈付けしたタグの数が減っている場合もある．また，注釈者5･6についても最後に行ったツール無しの場合には注釈付けしたタグの数が減っている．この点から考えると，注釈者の学習による影響が要因であるとは言えないと思われる．これは注釈事例の参照と，注釈付けされやすい部分の提示により，注釈者がより細かく注釈付けを行うことができたことを示している．

次に，注釈者間の注釈付けの一致について調べた．今回の実験では次に示す3つの場合を，注釈者間の注釈付けが一致したものと判定した．

最初に，完全一致である．これは全く同じ文字列に対して同一のタグが付与されている場合である．

次に，部分文字列一致である．これは完全一致ではないが，タグが付与されている両文字列に共通部分が存在する場合である．今回の実験では，形態素区切りや文節区切りを明確にして注釈付けを行ったわけではないために，タグの範囲についてはわずかに異なってしまう場合がある．

完全一致と部分文字列一致の例を図7に示す．

図7における部分文字列一致の例では，「強さ」の属性に対する属性値が「頑強」であるという点ではどちらの注釈者も注釈付けが一致している．しかし，程度を表す表現「それなりに」の部分をタグに含めているかどうかという点が異なっている．

```
完全一致：
注釈者A:<evaluation>優れています</evaluation>
注釈者B:<evaluation>優れています</evaluation>

部分文字列一致：
注釈者A:<value>それなりに頑強</value>
注釈者B: それなりに<value>頑強</value>
```

図7．完全一致と部分文字列一致の例

最後にattribute-valueタグの区切りの違いがある．これは，valueタグの注釈付け範囲の細かさの違いから起こるものである．例を図8に示す．

```
注釈者A:<attribute>操作</attribute><value>しにくい</value>
注釈者B:<value>操作しにくい</value>
```

図8．attribute-valueタグの区切りの違う例

図8の例について考えると，注釈者Bが省略されていると考えたattributeは「操作のしやすさ」である．また，注釈者Aは「操作」の表層部分が「操作しやすさ」を意味する属性であると認識している．この点を考えると，図8の場合は2人の注釈者が同一文字列に対して同じ認識をしていると考えられる．

また，図8の揺れは「サ変名詞+する」の組み合わせである点が揺れの理由として挙げられる．その間を区切るかどうかが注釈者によって分かれている．しかし，今回の実験ではこの点については注釈者に指示をしていなかった．今回の実験ではこのような場合はvalueについて一致しているものとして扱った．

上記の3つの場合を一致しているとし，タグ付与全体がどの程度の一致率であったかをF値で調べた．注釈者Aの注釈付けを正解とし，注釈者Bの注釈付けを評価するF値の計算方法は次のようになる．

$$\text{再現率} = \frac{\text{一致したタグ数}}{\text{注釈者Aの付与したタグ数}}$$

$$\text{適合率} = \frac{\text{一致したタグ数}}{\text{注釈者Bの付与したタグ数}}$$

$$\text{F値} = \frac{2 \times \text{再現率} \times \text{適合率}}{\text{再現率} + \text{適合率}}$$

今回の場合は，正解とする注釈者を2名のどちらとしてもF値は変わらない．

表6．注釈付け全体の一致率（F値）

	ツール無し	事例無し	事例有り
	1回目	2回目	3回目
注釈者1･2	0.61	0.59	*0.68*
	2回目	1回目	3回目
注釈者3･4	0.62	*0.66*	*0.66*
	3回目	2回目	1回目
注釈者5･6	0.64	0.68	*0.69*

表6を見てわかるように，どの場合においても，事例有りの場合が最も良い結果となった．注釈者が作業に慣れるに従って，一致度が上がっていくことも考えられるが，事例有りを最初に行った注釈者5･6の組においても事例有りの場合が最も優れた結果となった．これは，注釈付けにおける事例提示の有効性を示していると思われる．

次に，各被験者の組における注釈付けの一致率をタグごとに調べた．表7に結果を示す．

今回の結果では，多くの場合で事例有りの場合にκ値が高かった．一カ所だけ事例無しの結果の方がよくなっている部分があるが，これは事例有りの場合から先に実験を行ったグループでのことなので，注釈者が学習をしてしまった結果とも考えられる．

偶然一致していた可能性も否定できないが，事例有りから注釈付け作業を開始した注釈者の組であっても，他の注釈者の組と遜色ない一致率となっていることを考えても，注釈事例の参照は

有効であると思われる．また，事例有りから注釈付け作業を開始した注釈者の組は，同じデータに対して注釈事例の参照を行わなくなることで一致率が低下している．このことからも，注釈事例の参照は一致率を安定させるために有効であると言える．

表 7．注釈付けの一致率（κ値）

		ツール無し	事例無し	事例有り
		1 回目	2 回目	3 回目
注釈者 1･2	att	0.39	0.37	*0.54*
	val	0.51	0.59	*0.62*
	eva	0.63	0.35	*0.66*
		2 回目	1 回目	3 回目
注釈者 3･4	att	0.52	0.53	*0.54*
	val	0.52	0.56	*0.57*
	eva	0.57	0.67	*0.69*
		3 回目	2 回目	1 回目
注釈者 5･6	att	0.50	0.49	*0.52*
	val	0.56	*0.67*	0.64
	eval	0.58	0.55	*0.69*

4．3 注釈揺れの分析

本節では注釈揺れについて，細かく分析を行う．

注釈揺れには大きく分けて 2 種類が挙げられる．一つは，同一文字列に対して注釈をするかしないかの判断の違いである．もう一つは，同一文字列に対して異なるタグを付与する場合である．

前者について，一方の注釈者のみが注釈付けしたタグ数を表 8 に示す．

表 8．一方のみが注釈付けしたタグの数

		ツール無し	事例無し	事例有り
		1 回目	2 回目	3 回目
注釈者 1･2	att	*58*	34	52
	val	*77*	44	73
	eval	34	*49*	22
		2 回目	1 回目	3 回目
注釈者 3･4	att	56	39	*60*
	val	76	66	*81*
	eval	26	19	*28*
		3 回目	2 回目	1 回目
注釈者 5･6	att	19	25	*29*
	val	*56*	48	51
	eval	*25*	23	13

この表によると，実験の順番や，ツールの有無，注釈事例の有無の各条件と，一方の注釈者のみが注釈付けしたタグの数の間には相関が見られない．また，表 5 と比較してみても，付与されたタグの数が増えた場合に一方のみが注釈付けしたタグの数が増えているというわけではない．これは，注釈者間の認識の異なりよりも，どの程度の量の注釈付けを行おうとするかという注釈者各自の意識と注意力の違いが原因として考えられる．

次に，同一文字列に対して異なるタグが付与された場合を調べる．表 9 は同一文字列に対して異なるタグが付与された時のタグの違い方とその数を示している．

結果を見ると，value タグと evaluation タグの揺れが最も多かった．これは予備実験の結果と同じである．今回の実験では，この問題点についても必ずしも事例有りの場合に揺れが減っているとは言えない．さらに事例を増やす必要があると考えられる．

表 9．同一文字列に異なるタグが付与された数

		ツール無し	事例無し	事例あり
		1 回目	2 回目	3 回目
注釈者 1･2	val-eval	16	28	25
	val-att	1	4	4
	att-eval	0	3	1
		2 回目	1 回目	3 回目
注釈者 3･4	val-eval	21	20	18
	val-att	5	7	10
	att-eval	0	1	0
		3 回目	2 回目	1 回目
注釈者 5･6	val-eval	27	22	27
	val-att	1	6	0
	att-eval	0	5	0

さらに，上記の二つに当てはまらない揺れのなかには，注釈付け自体は異なるものの，注釈付けされた文字列の最後の文節については同じ注釈付けがされているものがあった．図 9 が例である．

注釈者 A：<evaluation>その日の気分によって聞けるというのがすばらしい</evaluation>
注釈者 B：<value>その日の気分によって聞ける</value>のが<evaluation>すばらしい</evaluation>

図 9．末尾部分が同じ注釈付けをされている例

図 9 の場合は注釈者 1 と注釈者 2 は共に「すばらしい」の部分は評価であると認識している．この揺れは，注釈者により注釈付けの粒度が異なるために発生すると考えられる．

また，注釈付けの粒度が異なるために起こる揺れが複数タグにまたがる場合があった．例を図 10 に示す．

注釈者 A:<attribute>茶葉</attribute>が<value>踊っているのが目に見える</value>
注釈者 B：<attribute>茶葉が踊っている</attribute>のが<value>目に見える</value>

図 10．複数タグにまたがる注釈揺れ

図 10 の例の場合，<value>タグのみに注目すると部分文字列一致となっている．しかし，前の<attribute>タグを見ると「踊っている」をどちらのタグに含めるかが異なっている．つまり，区切りが揺れている．

このような揺れの削減への対策の一つとして，注釈付けする区切りを，文節単位などに明確化することが考えられる．

5．おわりに

本稿では，人手による評判情報注釈付けに過去の注釈事例の参照を用いる手法提案した．また，注釈事例の参照を行う注釈付けツールを試作し，注釈付け実験を行った．結果として，注釈事例の参照を行った場合には複数注釈者間の注釈付けの一致率が向上した．

今後の課題としては，注釈者が value-evaluation の判断に迷う事例を蓄えること，より効果的な注釈事例提示方法の検討，事例として提示する文の類似度計算方の改良を行い，さらに詳細な評価実験を行うことが挙げられる．

その後，実際に複数注釈者の並列作業によるコーパス作成に取りかかる予定である．

参考文献

[1] 立石健二，石黒義英，福島俊一:インターネットからの評判情報検索，情報処理学会研究報告 NL144-11，pp.75-82 (2001)

[2] Peter Turney. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 417-424. (2002)

[3] 村野誠治，佐藤理史:文型パターンを用いた主観的評価文の自動抽出，言語処理学会第 9 会 年次大会発表論文集，pp.67-70 (2003)

[4] 飯田龍，小林のぞみ，乾健太郎，松本裕治，立石健二，福島俊一:意見抽出を目的とした機械学習による属性-評価値対同定，情報処理学会研究報告 NL165-4，pp.21-28 (2005)

[5] 峠泰成，大橋一輝，山本和英:ドメイン特徴語の自動取得による Web 掲示板からの意見文抽出，言語処理学会第 11 回年次大会発表論文集，pp.672-675 (2005)

[6] 洪陽杓，白井清昭:対話行為タグ付きコーパスの作成支援，言語処理学会第 11 会年次大会発表論文集，pp.815-818 (2005)

[7] 小林のぞみ，乾健太郎，松本祐治，立石健二，福島俊一:テキストマイニングによる評価表現の収集，情報処理学会研究報告 NL154-12，pp.77-84 (2003)

[8] 高村大也，乾考司，奥村学:極性反転に対応した評価値モデル，情報処理学会研究報告 NL168-22，pp.141-148 (2005)

[9] 小林のぞみ，乾健太郎，松本裕治:意見情報の抽出／構造化のタスク使用に関する考察，情報処理学会研究報告 NL171-18，pp.111-118 (2006)

[10] 宮崎林太郎，前田直人，森辰則:評判情報注釈付けタグセットの提案，言語処理学会第 12 回年次大会発表論文集，pp.240-243 (2006)

[11] 高橋哲朗，乾健太郎:アノテーションツール “Tagrin” の紹介，言語処理学会第 12 回年次大会発表論文集，pp.228-231 (2006)