意見を記述する表現のより深い意味解析

代表研究者　奥村　学　東京工業大学　精密工学研究所　教授
共同研究者　横野　光　東京工業大学　精密工学研究所　研究員
植田禎子　日本システムアプリケーション　研究員

1 要旨

我々人間が意見を表明する際に用いる表現は多様であり，字義通りの表現以外に，比喩や皮肉など，字義通り以上の意味を表す表現を用いることで，より深みのある意見を表明したりしている．このような背景から，これまでの意見分析に関する既存研究では全く扱われていない，比喩や皮肉など，意見を記述するのに用いられる「より深い」表現を計算機的に扱う技術を開発することを目指す.

2 背景

近年我々人間の周囲には，さまざまなメディアを通した情報が満ち溢れ，WWW上では，いわゆるマスメディアではない独自なメディアとして，掲示板(BBS)，チャット，ブログ(Weblog)などのような「口コミ」としての一般大衆による情報発信も盛んになりつつある．このような状況の中，一般の多くの人々が発信している大量の情報を有効に活用したいという要求も高まっている.一般大衆の発信する情報の中で特に関心を持たれているのは，一般大衆がどういう意見を持っているかに関する情報であろう．そのため,インターネット上の一般大衆の発信している意見を網羅的に収集，分析するシステムの研究開発が活発に進められている．我々はこれまでブログを定期的に監視し，そこから一般大衆の「生の声」と考えられる情報を抽出，発掘するためのシステムblogWatcherを開発している[1]．blogWatcherでは，収集したブログ集合にテキストマイニング技術を適用することにより，インターネット上の意見分析を実現している．しかし，我々の開発したシステムも含め，既存の意見分析エンジンはいずれも，字義通り評価を記述している表現を対象とし抽出，収集しているだけに過ぎない[2]．我々人間が意見を表明する際に用いる表現は多様であり，字義通りの表現以外に，比喩や皮肉など，字義通り以上の意味を表す表現を用いることで，より深みのある意見を表明したりしている.このような背景から,これまでの意見分析に関する既存研究では全く扱われていない，**比喩や皮肉など，意見を記述するのに用いられる「より深い」表現を計算機的に扱う技術を開発する**ことを目指す.

3 目的

本研究は，大きく次の2つの研究テーマを柱として進める.

1.　比喩や皮肉など，「より深い」表現を含む意見コーパスの作成，分析

比喩や皮肉などの表現を対象に構築された意見コーパスはこれまでにない.そこで，これらの表現を含む意見コーパスをまず収集するとともに，それらの表現を分析し，コーパスに必要なアノテーションを行う.

2.　比喩や皮肉など，「より深い」表現も扱える意見分析の計算モデルの開発

1.の分析で得られた知見および1.で構築したタグ付きコーパスを利用することで，比喩や皮肉などの表現も対象とした意見分析の計算モデルを開発する．研究方法欄で詳述するように，我々は，これらの表現を対象とした意見分析では，これらの表現が伝えたい字義通りの内容は何か(意見の極性を含む)，字義通りの

表現を用いる場合と比較してこれらの表現を用いることで得られる効果，の少なくとも2つを明らかにする必要があると考えている．

4 方法

以下各研究テーマについてそれぞれ研究計画，方法を述べる．

1. 比喩や皮肉など，「より深い」表現を含む意見コーパスの作成，分析

比喩や皮肉などの表現を含む意見コーパスをまず収集する．特定の対象に対する意見が現在，口コミサイトに大量にレビューテキストとして蓄積されている．そこで，口コミサイトに大量に蓄積されたレビューデータを意見コーパスとして収集し，それらの中の比喩や皮肉などの表現個所に着目するという手法を検討している．しかし，口コミサイトは本来，意見の収集，集約が目的となっているため，レビューテキストは比較的良質なテキストとして書かれる傾向があるとともに，特定の対象に対する意見のみが記述されるという性質があり，意見コーパスとして考えると，偏りのあるものとなっている懸念もある．そこで，レビューテキストとは別に，潜在的に意見を含むようなテキストとしてブログデータも別途収集し，レビューデータとの意見コーパスとしての性質の違いを分析してみたい．ブログ等，意見を収集，集約する目的のサイト以外で書かれている，潜在的に意見を含むようなテキストの場合，表現がよりくだけていて多様であり，多様な対象に対する意見が同一テキスト中に記述されているという性質を持っている．

次に，比喩や皮肉などの表現を分析し，コーパスに必要なアノテーションを行う．どのような観点から，これらの表現を分析し，どのようなアノテーションをコーパスに付与していくかも研究の過程で検討していく必要のある課題であると考えているが，現状では，少なくとも以下の点について着目する必要があると考えている．

a. これらの表現が伝えたい字義通りの意見の内容(対象のどういう属性についてどういう評価をしているのか)，その極性

たとえば，日本酒の香りについて「桃のような」という表現を用いる場合，「甘い香り」ということを伝えたいと書き手は考えており，肯定的な評価をしている．某ガムに対する意見としき手は考えており，否定的な評価をしている．

b. 字義通りの意見の内容がどの程度理解容易か(わかりやすいか)

「山のようなラーメン」は「量が多い」ことを伝えようとしていると理解できそうだが，「雪山のようなラーメン」はどのようなことを伝えようとしているのか，理解困難と言える．

c. 字義通りの表現を用いる場合と比較してこれらの表現を用いることで得られる効果

比喩や皮肉などの表現の効果は多岐にわたると考えており，分析の過程で整理，体系化していく必要があると考えているが，少なくとも現時点で以下のようなものが考えられる．

i. 具体的なイメージを想起させる

ii. 表現力の豊かさを感じさせる(たとえば，研究目的欄で例示した「ビロードのような」)

iii. 書き手，表現に対して肯定的/否定的な印象を与える(書き手，表現に対する評価)

「おいしい」という表現ばかり用いている書き手よりも，「かまどで炊いたようなご飯」という表現を用いる書き手の方が，ii.の効果の影響もあり，良い印象を与える傾向があるように思われる．

2. 比喩や皮肉など，「より深い」表現も扱える意見分析の計算モデルの開発

1.の分析で得られた知見および1.で構築したタグ付きコーパスを利用することで，比喩や皮肉などの表現も対象とした意見分析の計算モデルを開発する．上述したように，我々は，これらの表現を対象とした意見分析では，これらの表現が伝えたい字義通りの内容は何か，字義通りの表現を用いる場合と比較してこれらの表現を用いることで得られる効果，の少なくとも2つを明らかにする必要があると考えている．字義通りの内容をこれらの表現から同定する手法は，これまでの比喩，皮肉理解研究から知見が得られると考えてい

る．それらの知見を導入した計算モデルを構築し，1.a.の情報を自動的に表現から同定する手法を開発するとともに，その計算モデルにある種の確信度を出力させることにより，1.b.の情報が得られるようにできないか検討している．

5　結果

1.　比喩や皮肉など，「より深い」表現を含む意見コーパスの作成，分析

1.では，比喩や皮肉などの表現を含む意見コーパスをまず収集した．特定の対象に対する意見が現在，口コミサイトに大量にレビューテキストとして蓄積されている．そこで，口コミサイトに大量に蓄積されたレビューデータを意見コーパスとして収集し，それらの中の比喩や皮肉などの表現個所に着目するという手法を採った．具体的には，「AのようなB」の形式の名詞句を含む文を食べログのレビューテキストから収集し，以下の情報を付与してもらった．

- A と B の範囲
- ``A のような B''の極性
- 文の極性

極性とは，肯定的(positive)な評価なのか否定的(negative)な評価なのかを示す指標であり，``A のような B''の極性としては以下の 5 種類のうちの 1 つを付与した．

p〕表現単体で positive と判断できる
wp〕``A のような B''単体では判断できないが，周辺文脈から positive と考えられる
u〕極性無し
wn〕``A のような B''単体では判断できないが，周辺文脈から negative と考えられる
n〕表現単体で negative と判断できる

文の極性は``p''，``u''，``n''の 3 種類から 1 つを付与した．

事例を以下に示す．以下の事例は，CSV 形式で，前から順に``前半''，``のような''，``後半''，``コメント''，``A のような B の極性''，``文の極性''となっている．

```
″ビル内にある店舗なのに、そこは<A>リゾートホテル</A>″,″のような″,″<B>たたずまい</B>。″,″″,″p″,″p″
```

総事例数は 3000 件．そのうち，解析誤りなどで対象外となったものが 173 件あった．

「A のような B」という表現は，いくつかの種類に分類されることが分かっている．[3]では，推量（A と B が不明関係），比喩比況(A と B が不一致関係)，例示(A と B が包含関係)に分類している．

推量　ライオンのような動物がアフリカで取ったビデオに写っていた。
比喩比況　ライオンのような犬がいた。
例示　ライオンのような肉食動物は生肉からビタミン類を摂取する。

今回のデータでは，比喩比況の事例が多いので，それらの例を詳しく見ることにする．今回のデータに多く含まれていた表現は，<対象物 A>と<対象物 B>をある<類似点>で比況するような事例である．そして，この<類似点>が２つの対象物の属性となっている例が多く見られた．以下にいくつか例を挙げる．

```
2　ビル内にある店舗なのに、そこは<>リゾートホテル</>　　のような<>たたずまい</>。
6　磯っぺという小ぶりなパンも、同じくもちもちとした生地に、<>甘辛いみたらし</>　　のような<>味のたれ</>がかかっていて、海苔がのっています。

3　<>少年</>　　のような<>顔</>で歯を見せて
4　<>うどん</>　のような<>食感</>がないため、こちらで克服できた。
```

``A のような B''と文に振られた極性の分布を表 1 に示す．

表1　「AのようなB」表現の極性の分布

	p	wp	u	wn	n
``AのようなB''	301	179	2165	70	112
文	1138		1296		393

``A のような B''と文の極性の対応を表 2 に示す.

表2　「AのようなB」表現と文の極性の対応

文＼表現	p	wp	u	wn	n
p	280	140	674	22	22
u	8	9	1269	1	9
n	13	30	222	47	81

2.　比喩や皮肉など，「より深い」表現も扱える意見分析の計算モデルの開発

``AのようなB''の極性がPであるとき，AとBそれぞれに似ている語A', B'で構成される``A'のようなB'''の極性も同様にPであると考えられる．そこで，推定対象である``AのようなB''の極性として，その表現との類似度が最も 大きい``A'のようなB'''の極性を採用するモデルを考える.

``A のような B''と``A'のような B'''との類似度 sim(A,B,A',B')は以下の式で求める.

$$sim(A,B,A',B') = sim_{context}(A,A') * sim_{context}(B,B')$$

$sim_{context}(A,A')$は A と A'の文脈類似度を表す. 文脈類似度は ALAGIN から提供されている文脈類似語データベースから取得する． 学習データのどの語とも文脈類似度が計算できなかった場合は，推定に失敗したと考える.

「 A のような B」コーパスを用いて 10 分割交差検定を行った．結果を表 3 に示す．また，推定に失敗した事例の割合は 0.311 であった.

表 3　実験結果

極性	Precision	Recall	F1
p	0.582	0.343	0.432
wp	0.333	0.135	0.192
u	0.880	0.654	0.750
wn	0.379	0.157	0.222
n	0.298	0.155	0.204

6　今後の課題

残念ながら，予算の関係で比較的小規模なデータしか作成することができず，また，意見分析の計算モデルも素朴なものしか開発できなかった．今後データをさらに拡充するとともに，より洗練された計算モデルを開発していく予定である.

【参考文献】

［1］奥村 学，南野 朋之，藤木 稔明，鈴木 泰裕，日本語blogページの自動収集と監視に基づくテキストマイニング，情報科学技術レターズ，pp.71-73，2004.

［2］大塚裕子，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，2007.

［3］森山卓郎，推量・比喩比況・例示ー「よう／みたい」の多義性をめぐってー，宮地裕・敦子先生 古稀記念論集 日本語の研究，明治書院，1995.

〈発　表　資　料〉

題　　名	掲載誌・学会名等	発表年月