blogマイニングと評判分析

※本発表では「評判分析」を含む
「評価分析」を紹介します．

奥村 学
東京工業大学
乾 孝司
JSPS

Agenda

- 準備
  - ◆評価分析とは？
  - ◆応用／題材／歴史
- 評価分析の要素技術

評価分析とは？

- ある対象の評価を記述しているテキスト断片に対して，その評価極性（肯定的な評価 or 否定的な評価）を推定すること

レストランAは味がよい	肯定
喫茶Bのコーヒーはまずい	否定

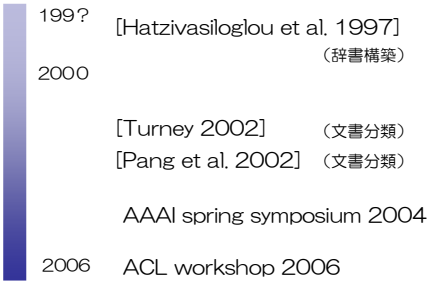
評価分析の応用

- blogマイニング
- マーケティング／リスク管理（企業）
- 商品購入時の判断材料（ユーザ）

評価分析の題材

- 意見の収集，集約が目的となっているもの
  - ◆自由回答アンケート，レビューサイト
  - ◆カスタマーサポートセンター「お客様の声」
  - ❖比較的良質な文書，話題が限定的
- 潜在的に意見を含むもの
  - ◆blog，Web掲示板，チャット
  - ❖くだけた表現が多様，話題が雑多

評価分析研究の歴史



評価分析の要素技術

評価分析の要素技術

■３つの代表的な技術

	扱う粒度
◆評価表現辞書の構築	単語
◆評価情報を観点とした文書分類	文書
◆評価情報の要素組の抽出と分類	単語の組合せ

⬆
blogマイニングとの関連が強い

評価分析の要素技術

■評価表現辞書
- ◆評価表現とその評価極性のペア集合
  - ❖文脈独立

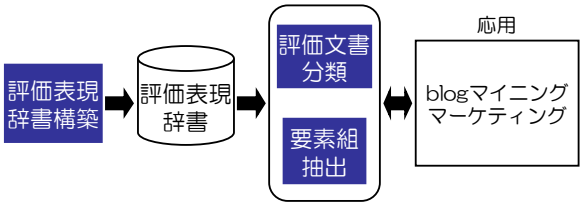
良い ― 肯定　美味しい ― 肯定
悪い ― 否定　まずい ― 否定

■評価情報の要素組
- ◆評価表現＋評価対象，属性など
  - ❖文脈依存

店Cのコーヒーは美味しいと思う ➡ <店Cのコーヒー，美味しい> 肯定
店Dのコーヒーは美味しくない ➡ <店Dのコーヒー，美味しくない> 否定

評価分析の要素技術

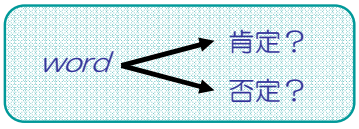
■要素技術の関係



評価表現辞書の構築

評価表現辞書の構築

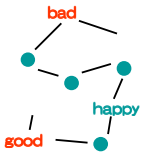
■単語を肯定極性と否定極性に分類
- ◆特に，形容詞が重要



- ◆語彙ネットワークを利用した手法
- ◆共起情報を利用した手法

評価表現辞書の構築
語彙ネットワークを利用した手法

- [Kamps et al. 2004]
  - ◆類義関係にある語の評価極性は一致しやすい
  - ◆WordNet　形容詞のsynonymy
  - ◆“good”，“bad”との近さに注目



$$\frac{d(\text{word}, \text{bad}) - d(\text{word}, \text{good})}{d(\text{good}, \text{bad})}$$

d：2つの形容詞間の最短経路長

[Kamps et al. 2004] Jaap Kamps, Maarten Marx, Robert J. Mokken and Maarten de Rijke. Using WordNet to Measure Semantic Orientations of Adjectives. LREC2004.

評価表現辞書の構築
共起情報を利用した手法

- [Turney 2002]
  - ◆肯定（否定）極性語の周辺には肯定（否定）極性語が現れやすい
  - ◆コーパスから共起情報を獲得
  - ◆“excellent”，“poor”のどちらと共起しやすいか

$$PMI(\text{word}, \text{excellent}) - PMI(\text{word}, \text{poor})$$

$$PMI(a,b) = \log \frac{p(a,b)}{p(a)\,p(b)}$$

[Turney 2002] Peter D. Turney. Thumbs up? thumbs down? semantic orientation applied to unsupervised classification of reviews. ACL2002.

評価表現辞書の構築
語彙ネットワークを利用した手法

- [Kamps et al. 2004]
  - ◆形容詞以外
  - ◆WordNetにエントリのない語　に対応できない
- [Turney 2002]
  - ◆任意の語について計算可能，ただし
  - ◆大規模な共起データの簡易な入手方法が必要

評価表現辞書の構築

- [Hatzivasiloglou et al. 1997]
  Vasileios Hatzivassiloglou and Kathleen R. McKeown. Predicting the Semantic Orientation of Adjectives. ACL1997.
  - ◆コーパス中の「形容詞 – 接続詞 – 形容詞」
- [那須川ら 2004]
  那須川 哲哉, 金山 博. 文脈一貫性を利用した極性付評価表現の語彙獲得. 情報処理学会自然言語処理研究会(NL-162-16), 2004.
  - ◆文脈中での評価極性の一貫性
- [Takamura et al. 2005]
  Hiroya Takamura, Takashi Inui and Manabu Okumura. Extracting Semantic Orientation of Words using Spin Model. ACL2005.
  - ◆Spin glass model

評価表現辞書の構築

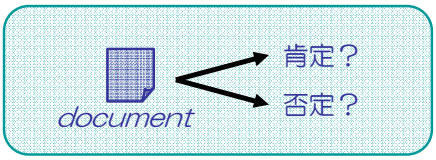
- 辞書構築の評価
  - ◆The General Inquirer [Stone et al. 1966]
  - ◆http://www.wjh.harvard.edu/~inquirer
  - ◆テキスト内容分析のための言語知識データ
  - ◆“Positiv” ラベルと“Negativ”ラベル
  - ◆英語

[Stone et al. 1966] P. J. Stone and D. C. Dunphy and M. S. Smith and D. M. Ogilvie. The General Inquirer: A Computer Approach to Content Analysis. MIT Press. 1966.

評価情報を観点とした文書分類

評価情報を観点とした文書分類

■文書を肯定極性／否定極性に分類



◆教師あり学習に基づく手法
◆評価情報の比率に基づく手法

評価情報を観点とした文書分類
教師あり学習に基づく手法

■[Pang et al. 2002]
◆映画レビューを肯定／否定に分類
◆ナイーブベイズ，最大エントロピー法，ＳＶＭｓ
◆単語uni-gram，単語bi-gram
◆ＳＶＭｓ＋単語uni-gram：精度82.9%

[Pang et al. 2002] Bo Pang, Lillian Lee and Shivakumar Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. EMNLP2002.

評価情報を観点とした文書分類
教師あり学習に基づく手法

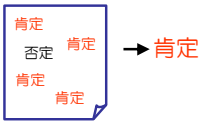
■[Mullen et al. 2004]
Tony Mullen and Nigel Collier. Sentiment analysis using support vector machines with diverse information sources. ACL2004.
◆評価表現を利用

■[Pang et al. 2004]
Bo Pang and Lillian Lee. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. ACL2004.
◆意見をあらわす文に注目

■[Matsumoto et al. 2005]
Shotaro Matsumoto, Hiroya Takamura and Manabu Okumura. Sentiment Classification using Word Sub-Sequences and Dependency Sub-Trees. PAKDD2005.
◆語の系列，依存木を利用

評価情報を観点とした文書分類
評価情報の比率に基づく手法

■[Turney 2002]
◆文書の評価極性は文書内の評価表現のみから決定される
◆文書中の評価表現がもつ評価極性の平均値に従って肯定／否定に分類
◆精度65.8%～84.0%



PMI(word, excellent) － PMI(word, poor)

評価情報を観点とした文書分類
評価情報の比率に基づく手法

■[Taboada et al. 2004]
Maite Taboada and Jack Grieve. Analyzing Appraisal Automatically. AAAI-EAAT2004.
◆書き手の主要な意見は文書全体に均等に現れるのではなく，特定の部分に集中している
◆評価表現の出現位置による重みづけ

■[Kennedy et al. 2005]
Alistair Kennedy and Diana Inkpen. Sentiment classification of movie and product reviews using contextual valence shifters. FINEXIN2005.
◆極性変化子（contextual valence shifter）
◆“not good”，“very good”

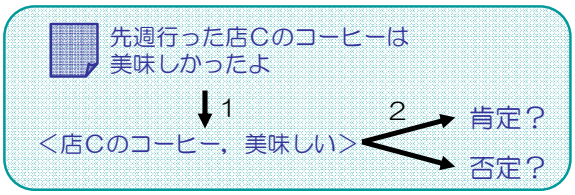
評価情報を観点とした文書分類

■評価情報を観点とした文書分類の評価
◆映画レビューデータ[Pang et al. 2002]
◆http://www.cs.cornell.edu/people/pabo/movie-review-data/
◆英語

評価情報の要素組の抽出と分類

評価情報の要素組の抽出と分類

■要素組を肯定極性／否定極性に分類



Step１：要素組の抽出
Step２：要素組の評価極性を分類

評価情報の要素組の抽出と分類

■文脈を考慮して評価極性を分類

この りんご は 美味しい	→	肯定
この りんご は 美味しく ない	→	否定
この りんご は 美味しかっ た？	→	評価なし
この ベッド は 眠気を誘う	→	肯定
この 講義 は 眠気を誘う	→	否定

評価情報の要素組の抽出と分類

■要素組の抽出（要素の特定＋要素の関連づけ）
- ◆評価情報の要素　<りんご，酸味，素敵だ>
  - ■評価表現（素敵だ）　評価表現辞書
  - ■評価対象（りんご）　分析対象として与えられる
  - ■属性（酸味）　自動処理
- ◆属性辞書構築
  - ■対象ごとに用意
  - ■パターン[小林ら2005]，統計量[Yi et al. 2005]
- ◆要素の関連づけ
  - ■構文情報に基づく素朴な手法

評価情報の要素組の抽出と分類

■要素組の評価極性を分類
■[鈴木ら2004] in BlogWatcher

[鈴木ら2004]鈴木泰裕, 高村大也, 奥村学. Weblogを対象とした評価表現抽出. 人工知能学会セマンティックウェブとオントロジー研究会(SW-ONT-A401-02), 2004.

[小林ら2005]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, 2005.

[Yi et al. 2005]Jeonghee Yi, Wayne Niblack. Sentiment Mining in WebFountain. ICDE2005.

その他

■意見性：意見か否か
- ◆[Wiebe et al. 2004]
  J. Wiebe, T. Wilson, R. Bruce, M. Bell and M. Martin. Learning subjective language. Computational Linguistics, Vol.30, No.3, 2004.

■肯定／否定からの拡張
- ◆neutral [Koppel et al. 2005]
  Moshe Koppel, Jonathan Schler. The importance of neutral examples for learning sentiment. FINEXIN2005.
- ◆five star review[Pang et al. 2005]
  Bo Pang, Lillian Lee. Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales. ACL2005.

まとめ

- 評価分析の要素技術
  - ◆評価表現辞書の構築
  - ◆評価情報を観点とした文書分類
  - ◆評価情報の要素組の抽出と分類

参考文献

- AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications (AAAI-EAAT), 2004.
- Sentiment and Subjectivity in Text, Workshop at ACL2006.
- 乾孝司, 奥村学.
  テキストを対象とした評価情報の分析に関する研究動向.
  自然言語処理, Vol.13, No.3, 2006.（掲載予定）

上記論文に参考文献リストを掲載しています.

宣伝

- 「感情・評価・態度と言語」ワークショップもお楽しみに…
- 「自然言語処理」特集号「感情・評価・態度と言語」にも多数のご投稿をお待ちしています.
- 「意見分析エンジン」，大塚，乾，奥村，コロナ社，近刊.