スケーラブルな評価表現解析 …を目指してます

東京大学　生産技術研究所
鍛治伸裕

はじめに

- やりたいこと
  - ある対象の評価が記述されたテキスト（評価テキスト）を解析して，その評価内容を整理して提示する技術の開発．
  - マーケティングへの利用，企業の風評の監視などへ応用．
- これまでの技術
  - 自由回答アンケートやレビューサイトなど，限定されたテキストが対象となっている技術．
- 目指している技術
  - ウェブ上のあらゆるテキスト（とくにブログや掲示板）を処理できるようにスケールアップされている技術．

具体的に評価表現解析とは

- 文や語句の評価極性判定
  - 肯定的か否定的かの二値分類（中立カテゴリを導入することも）
  - 例えば
    - XXXはソフトウェアのバージョンアップが容易で安価 （肯定的）
    - ～すると，XXXは実行速度が落ちる （否定的）
- 評価が記述されたテキストの構造化
  - 〈対象，属性，評価値〉の三つ組みを抽出
  - 上の例文の場合
    - 〈XXX，バージョンアップ，容易〉
    - 〈XXX，（値段），安価〉
    - 〈XXX，実行速度，落ちる〉

研究としての狙いどころは？

Q1： 対象テキストを限定しない評価表現解析は，技術的に何が難しいのか？

Q2： 単に処理テキストが増えただけ？計算量だけの問題なのでは？

A： 単に計算量の問題だけではない．問題設定が変わると，以下の二点が技術的に難しくなる．

1. 解析に必要な知識の獲得
2. 評価テキストの取得

難しさ（1）： 知識獲得

- 評価表現解析には“知識”が必要
  - 評価表現辞書
  - 評価極性タグ付きコーパス（テキスト中の評価表現がアノテートされている）
- これまでの方法は
  - 所与ドメインの小規模な辞書/コーパスを人手で作成
  - レビューテキストをタグ付きコーパスとして利用（ノイズが多いの使うのは難しいが…）

知識獲得の問題点

- 対象テキストを限定しない場合，以下の問題が顕在化するだろう

1. 規模
   - 大規模なテキストを相手にするには，大規模な知識が必要
   - 人手で作成するのはコストが高い（鈴木ら，2004）
2. ドメイン
   - 幅広いドメインをカバーした知識が必要（ドメインごとに準備するのは現実的でない）
   - あるドメイン特有の知識は他ドメインに適用できない(Aue and Gamon, 2005)

難しさ(2)： 評価テキストの取得

- 所与キーワードの評価テキストが自由に手に入らないと始まらない
  - アンケートやレビューを対象としていれば問題にならない
  - ほとんど議論されていない(乾 and 奥村, 2006)
- 例えば「tsubaki」の評判を調べたいとき
  - 資生堂が2006年春から展開しているヘアケアブランドの名前である．資生堂のメガブランド構想第4弾．キャッチコピーは「日本の女性は美しい」（wikipedia から引用）
  - 単純に「tsubaki」で検索してみると…

某大手スーパーのシャンプー売り場で「本日販売開始」という文字につられて見てみると、カネボウから新発売されたシャンプー・リンスが。
その名も「いち髪」

「いち髪」という「tsubaki」の競合を買った．

とりあえず、限定トライアルセットを購入。

「いち髪」のCMについて語っている．
ここで「tsubaki」の名前が出てくる．

イメージキャラクターは深津絵里のようだが、[illegible]の。
先ほど娘が、いち髪のCMやってたよ～～と言っていたがすでに遅し。残念！
言葉の由来は、一髪二姿（いちかみにすがた）からきているみたい。
花王「アジエンス」資生堂「TSUBAKI」に対抗しているのでしょう。

早速今晩使ってみましたが、洗っている時からのゴワゴワ感もなくいい感じです。
山桜の香りもきつくなく、とてもいい香り。
これは好きな香りなのでお気に入りで～す。

「いち髪」の感想．
結局「tsubaki」に関する言及はなし．

これまでの成果

- こうした問題を一度に全て解くのは難しい
- 手始めとして，評価文の自動収集手法を考案した
  - 広いドメインのテキストを大量に集める
- 収集された評価文の利用方法として
  - 文の極性判定の訓練データなどに使う
  - ここから評価表現を抽出する，ということを考えている

（例）

（肯定的） ベッドを動かして掃除する手間が省ける．
（否定的） 一般道での燃費が少し悪い．

Automatic Construction of Polarity-tagged Corpus from HTML Documents

Nobuhiro Kaji and Masaru Kitsuregawa
COLING/ACL 2006, Poster Sessions

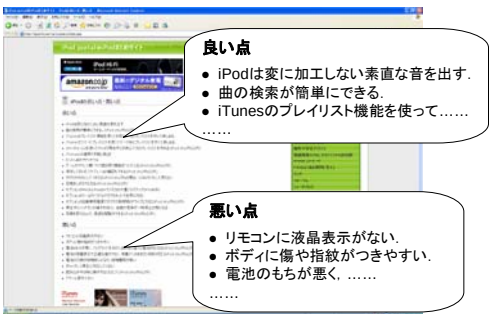
大規模ウェブテキストを利用

- テキストからの知識獲得は古典的なテーマ
  - 上位下位関係の獲得 (Hearst92)

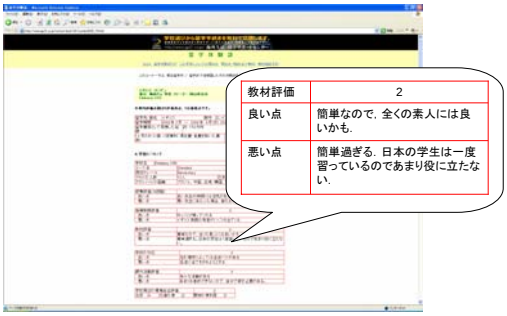
… works by *such* authors *as* Herrick, Goldsmith …
… countries *including* Canada and England …

- “大規模な”ウェブテキストの利用は一つのトレンド
  - Google 5-gram(1,011,582,453,213 words)
  - 日本語格フレーム辞書(5億文)

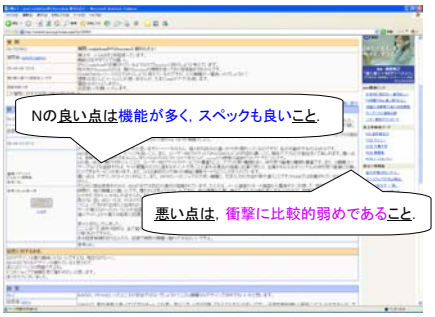
1．箇条書き形式



2. 表形式

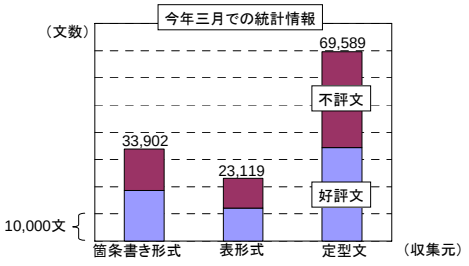


3. 定型文



コーパス構築

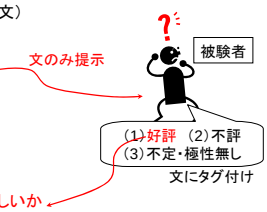
- 約9億件のHTML文書から約50万の評価文を収集
  - 公開URL: http://www.tkl.iis.u-tokyo.ac.jp/~kaji/acp
  - コーパスは現在も継続的に収集中



主観評価

500文を無作為抽出（好評259文，不評241文）

極性	評価文
好評	何と言っても，料金が良心的だ．
好評	順応性が素晴らしくある．
不評	エンジンが非力で少々うるさい．
	……

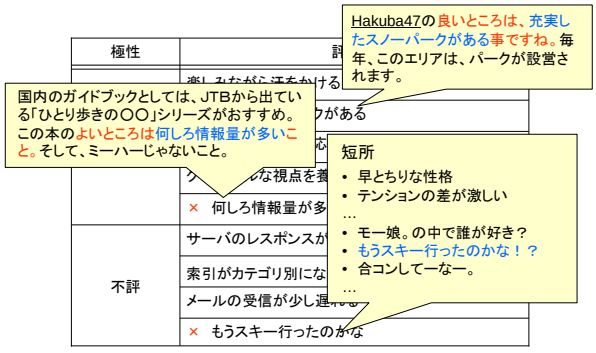


二人の被験者による結果

	正解率
被験者A	459/500 (91.5%)
被験者B	460/500 (92.0%)

被験者間の一致率 = 467/500 (93.4%)
Kappa 値 = 0.90

具体例



議論

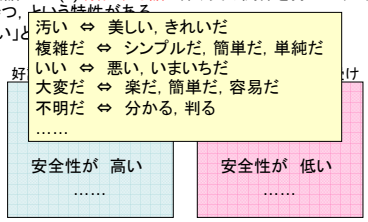
- コーパスの規模
  - 既存のものと比較して十分大規模．
  - しかし，この規模でもデータスパースネスの問題は深刻であると考えている．今後は，(1) より効率の良いコーパス収集法 (2) 言い換え技術の適用を検討中．
- 好不評タグの精度
  - 主観評価で90%強（曖昧な文を誤りにカウント）は，実用的なレベルと考えている．
  - 評価文分類実験でも良好な結果を得た．

議論

- 好不評のバランス
  - 今回報告した実験結果では，好評も不評もほぼ同数．
  - ただし，色々なパターンを追加すると不評が多くなるケースも見られた．
- ドメイン
  - ドメイン横断的なコーパス獲得を行った．
  - ドメインによって評価極性が変化する表現は，原理的にうまく扱うことができない（例えば「くだらない」など）．
  - もし，上記のような表現が多い場合は，ドメイン適用処理やドメイン知識の追加を検討したい．

議論

- 本コーパスに意外な利用法はないか？
  - もともと評価表現を獲得するために作ったコーパスだが
- 例えば，統計的に反義語を学習できるのでは
  - 反義語には (1) 類似する語と係り受け関係を持ちやすく，(2) 逆極性を持つ，という特性がある
  - 「高い」と



まとめ

- “あらゆるテキスト”を対象とした評価表現解析とその難しさ
- 評価文の自動収集法の提案
- 今後の展望
  - 評価テキストの取得方法の考案
  - 評価表現の解析結果をマーケティングに活用
  - 会社の風評監視などへの適用の検討