CSE 6412 – Data Mining
Research Paper Presentation

Opinion Mining in Comparative Sentences

by Murthy Ganapathibhotla, Bing Liu
(2008)

Presenter: Nadine Dulisch

1. Introduction
2. Problem Definition
3. Proposed Technique
4. Identifying Preferred Entities: The Algorithm
5. Evaluation

- Objective:
  Mining opinions from comparative sentences, i.e. which entities in a comparison are prefered by its author
- Distinction to direct opinion sentences:
  - “the picture quality of Camera X is great”
  - “the picture quality of Camera X is better than that of Camera Y”
    - 2 entities compared (Camera X, Camera Y)
    - Shared features or attributes (picture quality)
  - About 10% of sentences are comparisons (of user generated content)

- Jindal and Liu (2006) proposed technique to identify comparative sentences
  - Extract entities, comparative words, entity features
  - *“Camera X has a longer battery life than Camera Y”*
    - Entities: Camera X, Camera Y
    - Comparative word: longer
    - Feature: battery life
  - Doesn't determine, which entity is preferred

  ⇒ Objective of this paper

- Basic information unit: sentence
- Sentences usually contain
  - a comparative word (better, worse, -er word) or
  - a superlative word (best, worst, -est word)

  (here: “comparative word” used for both)
- Entities often appear on both sides of a comparative word
- Preferred entity mainly determined by comparative word

- Opinionated comparative words:
  - Explicitely indicate user preferences (better, worse, best)
  - “the picture quality of Camera X is ***better*** than that of Camera Y”
- Context-dependent opinion comparatives:
  - Not opinionated comparative words; opinion orientations depend on context or application domain
  - “longer” in itself is not opinionated
    (length of a feature of an entity is greater than the length of the same feature of another entity)
  - can be a desired or undesired state
    (e.g. camera: longer battery life (positive), program: longer execution time (negative))

=> implicit opinion

- Opinionated comparative words usually easy to handle
- Problem: How to identify opinion orientations?

  1) What is the context?
  - the whole sentence (but that's too complex)
  - Smallest context, that can determine the orientation:
    - Entity features being compared
    - comparative word

2) How to use the context to determine the orientation?

- External information/knowledge needed
  - For this paper: Customer reviews on the web (epinions.com)
  - Seperated pros and cons (so positive and negative opinions are known)
  - Drawback: pros and cons seldom contain comparative words

    => compute whether the comparative word and the feature are more associated in Pros than Cons

- Entity
  - An entity is the name of a person, a product, a company, a location, etc, under comparison in a comparative sentence.
- Feature
  - A feature is a part or attribute of the entity that is being compared.
- “Camera X’s battery life is longer than that of Camera Y”
  - Entities: “Camera X” and “Camera Y”
  - Feature: “battery life”

- Types of comparatives:
  - Non-equal gradable
    - Relations that express **total ordering** of some entities with regard to their shared features (“Camera X's battery life is longer than that of Camera Y”)
  - Equative
    - Relations that state 2 objects as **equal** with respect to some features (“Camera X and Camera Y are about the same size”)
  - Superlative
    - Relations that rank **one object over all others** (“Camera X's battery life is the longest”)
  - Non-gradable
    - Sentences, which compare features of two or more entities, but don't explecitely grade them (“Camera X and Camera Y have different features”)

- Comparative Relation
  - <ComparativeWord, Features, EntityS1, EntityS2, Type>
  - ComparativeWord: Keyword to express comparative relation in the sentence
  - Features: Set of features being compared
  - EntityS1, EntityS2: Sets of entities being compared
  - Type: non-equal gradable, equative, superlative
  - Example:
    - “Camera X has longer battery life than Camera Y”
    - <longer, {battery life}, {Camera X}, {Camera Y}, non-equal gradable>

- Assumption
  - The work in (Jindal and Liu 2006) has extracted the comparative relation from a comparative sentence
- Objective
  - Given the extracted relation, identify whether the entities in EntityS1 or EntityS2 are preferred by the author

- Comparatives and superlatives special forms of adjectives and adverbs
  - Regular comparatives
    - Type 1 comparatives (and superlatives)
      - suffixes “-er”, “-est”
    - Type 2 comparatives
      - Adjectives and adverbs with 2 or more syllables
      - Formed with “more”, “most”, “less”, “least” (e.g. more beautiful)
  - Irregular comparatives
    - More, most, less, least, better, best, worse, worst, further/farther, furthest/farthest
    - Grouped under Type 1, because they behave similarly

- Non-standard words that express gradable comparisons
  - Prefer, superior
  - “in term of battery life, Camera X is superior to Camera Y” => Camera X is preferred
  - Grouped under Type 1, since they behave similarly
  - For this paper, list of 27 words obtained from (Jindal and Liu 2006)

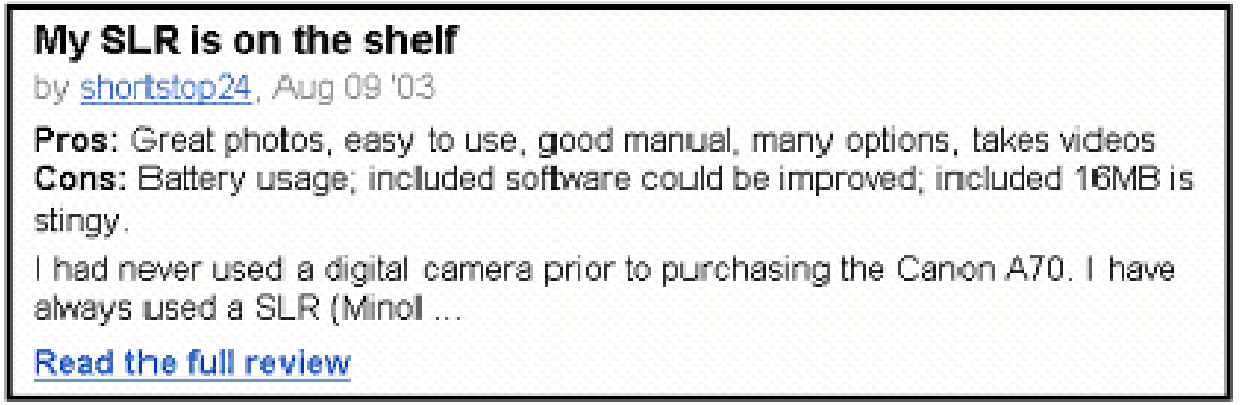
- Categories of comparatives in regards to increasing/decreasing value
  - Increasing comparatives
    - Expresses an increased value of quantity (more, longer)
  - Decreasing comparatives
    - Expresses a decreased value of quantity (less, fewer)
- Categories of comparatives in regards to whether they carry positive or negative sentiments/opinions
  - Opinionated comparatives
    - Type 1: better, worse (explicit opinion)

- Type 2: more, less, most, least added; opinion determined by both words
  - "increasing comparative" Negative → Negative Opinion
  - "increasing comparative" Positive → Positive Opinion
  - "decreasing comparative" Negative → Positive Opinion
  - "decreasing comparative" Positive → Negative Opinion

- Comparatives with context-dependent opinions
  - Used to compare gradable quantities of entities (e.g. Type 1: higher, lower)
  - "Car X has higher mileage per gallon than Car Y"
  - Domain knowledge needed to know, whether higher is positive or negative
  - Type 2: similar to rules above (comparative word (more, less), adjective/adverb and feature important)

- Approach in this work
  - Usage of opinion word list from (Hu and Liu 2004)
  - For opinionated comparatives, conversion of adjectives/adverbs (in the list) to their comparative forms
    - automatically, based on grammar rules and WordNet (large lexical database of English)
    - if a word is positive, then also their comparative/superlativ form
  - Manual categorization into increasing and decreasing comparatives

- Conjecture: context formed by comparative and feature
- For type 2 comparatives: feature and adjective/adverb
  - “Program X runs more quickly than Program Y”
  - (“run”, “quickly”)
    - If positive (based on external information) =>conclusion that “Program X” is preferred (since “more” is an increasing comparative)
- Contexts used to find opinion orientations of comparatives



My SLR is on the shelf

by shortstop24, Aug 09 '03

Pros: Great photos, easy to use, good manual, many options, takes videos
Cons: Battery usage; included software could be improved; included 16MB is stingy.

I had never used a digital camera prior to purchasing the Canon A70. I have always used a SLR (Minol ...

Read the full review

- Each phrase/sentence segment usually contains an entity feature and an opinion word
- User opinion on feature is clear (positive, negative)
- Separation using punctionation and words (e.g. and, but)

great photos	<photo>
easy to use	<use>
good manual	<manual>
many options	<option>
takes videos	<video>

- Preparatory work:
  - Usage of Pros and Cons to determine whether a comparative/entity feature combination is positiv or negativ
  - Comparatives seldom used in Pros and Cons
  - To find comperatives and entity features, convert comparatives to their base form
    (automatically with WordNet and grammar rules)
- Putting everything together to identify the preferred entity:
  - C = comparative word, F = feature
  - 2 main cases for the 2 types of comparatives
    (Type 1: suffixes “-er”, “-est”, etc; Type 2: formed with “more”, “less”, etc.)

- Case 1: Type 1 Comperative or Superlative
- 4 sub-cases

1.A.) C is opinionated

- ```
  If C has positive orientation then
     preferred entity = EntityS1
  else
     preferred entity = EntityS2
  ```

- Assignment is temporarily, because sentence may contain negations (e.g. “not”)
  ```

1.B.) C is not opinionated, but F is opinionated

- “Car X generates more noise than Car Y”
- “noise” is a negative noun
-
```
If orientation of F = positiv and C is
increasing comparative word then
   preferred entity = EntityS1
else
   preferred entity = EntityS2
```

- All possibilities/rules:
  “increasing C” + Positive → EntityS1 preferred
  “decreasing C” + Positive → EntityS2 preferred
  “increasing C” + Negative → EntityS2 preferred
  “decreasing C” + Negative → EntityS1 preferred

1.C.) Both F and C are not opinionated

- External information needed (Pros and Cons from Reviews)
- Look for F and C in (i.e. the context) in list of phrases in Pros and Cons
- Find whether combination is positive or negative
  - Compute their associations in Pros and Cons
  - If more associated in Pros than Cons => positive sentiment
- Association measures:
  - Point-wise mutual information (PMI)
  - One-side association (OSA) (proposed in this paper)

- PMI

$$PMI(F,C) = log \frac{Pr(F,C)}{Pr(F)Pr(C)}.$$

  - Commonly used for computing association of 2 terms
  - Not suitable for this problem
    - PMI is symmetric (PMI(F,C) = PMI(C, F))
    - F and C not symmetric
      - Feature usually modified by particular adjective word
      - But the adjective word can modify several features (long battery life, long execution time)
    - => probability of C given F (Pr(C | F)), confidence in data mining
      - Not suitable when C occurs frequently and F rarely (high probability may just represent pure chance, association may be spurious)
    - => OSA

- OSA

$$OSA(F,C) = log \frac{Pr(F,C)Pr(C|F)}{Pr(F)Pr(C)}$$

  - Pr(C | F) biases the mutual association of F and C to one side

  - Computation for positive and negative ($OSA_P(F, C)$; $OSA_N(F, C)$)
  - Decision Rule:
    - If $OSA_P(F, C) - OSA_N(F, C) \geq 0$ then

```
      EntityS1 is preferred
   else
      EntityS2 is preferred
```

- Computing $OSA_P(F, C)$
  - For C additionally use base form, synonyms, antonyms (obtained from WordNet, currently only single word features); for F additionally use synonyms
  - $Pr_P(F, C)$
    - + 1 for every time, C and F co-occur in Pros phrase
    - + 1 for every time, if antonym of C and F co-occur in a Cons phrase
    - Usage of both Pros and Cons allows to find more occurances (to produce more reliable results)
  - $Pr_P(F)$; $Pr_P(C)$
    - C: synonyms in Pros, antonyms in Cons considered
    - F: synonyms in Pros and Cons
- Computing $OSA_N(F, C)$ accordingly

1.D.) C is a feature indicator

- The feature doesn't appear explicitely in the sentence, but is implied
- Words that imply the feature = feature indicator
- “Camera X is smaller than Camera Y”
- “smaller” feature indicator for feature “size”
- # of times C occurs in Pros/Cons (n+/n-)
-
```
If n+ ≥ n- then
   EntityS1 preferred
else
   EntityS2 preferred
```

- Case 2: Type 2 Comperative or Superlative
- 2 sub-cases
  2.A.) Adjective/adverb in the comparison is opinionated
  - In this case, feature F is not important
  - “Car X has more beautiful interior than Car Y”
  - more = increasing comparative, beautiful = adjective with positive orientation, (interior = feature)
  - Car X clearly preferred
  - Handled similar to 1.B. (C is not opinionated, bug F is opinionated)

2.B.) Adjective/adverb in the comparison is not opinionated

- If adjective is a feature indicator => handled according to 1.D.
- Otherwise, form context using the feature and the adjective/adverb => handled according to 1.C.
- Result combined with comparative word before the adjective to decide based on rules in 1.B

- **Negations:**
  - The other entity is preferred
  - Can be problematic (“not longer” doesn't mean “shorter”)

- System PCS (Preferred entities in Comparative Sentences) has been implemented
  - No system to compare with exists
- Evaluation Datasets
  - Consists of 2 subsets
    - (Jindal and Liu 2006): product review and forum discussion sentences (digital cameras, DVD players, Intel vs AMD, etc)
    - Self collected data: forum discussion data about mobile phones and reviews from amazon.com and cnet.com
  - 837 comparative sentences, collected from thousands of sentences (about 10 % comparative sentences)
  - 84% of comparative sentences had EntityS1 as preferred entity => people tend to put the preferred entity first
  - Each 15162 Pros and Cons extracted (epinions.com)

- Results
  - If just EntityS1 is anounced as preferred entity: 84% accuracy
  - PCS using OSA measure: 94.4% accuracy
  - Mainly precision, recall and F-score used for evaluation
  - Better results if EntityS1 is preferred
    - Observation: sentences tend to be more complex, if EntityS2 is preferred
  - Comparison without the use of Pros and Cons:
    - When context dependency handling is required, take majority as default (EntityS1)
    - Precision improvement of PCS statistically significant at 95% confidence level

- Looking at the 187 sentences, that need dependency handling
  - 72.2% have EntityS1 as preferred entity
  - PCS reaches 89.6% precision for EntityS1 as preferred entity (69.6% for EntityS2)
- OSA vs PMI
  - OSA better in F-score (1.2% for EntityS1, 2% for EntityS2)
  - OSA gains for whole dataset are less, because number of sentences that need context dependency handling is small (22%)

Thank you for your attention.

Any Questions?