Opinion Extraction based on Syntactic Pieces

Suguru Aoki and Kazuhide Yamamoto
Nagaoka University of Technology
Introduction

- Opinion extraction is
  - To find out personal opinions
    - Such as reputation and dissatisfaction with product, service, and so on
  - Many works
    - Decide semantic orientation (positive or negative) to word, phrase or etc., and
    - Classifying opinion sentence (document) or not.
Related Works

- Document classification
  - Fujimura et al. (2004)
    - Using bag-of-words, noun/adjective/adjectival verb
    - Semantic orientation of word changes with domains
    - Longer processing unit necessary
  - Turney (2002)
    - Using Adjective phrase, such as n-gram
    - N-gram does not work well for agglutinative languages
    - Some kinds of syntax should be required
Related Works

- Opinion expression extraction
  - Tateishi et al. (2004)
    - Using Opinion triplet : \{object, attribute, evaluation\}
    - Make a triplet dictionary
    - Extract only defined patterns, therefore few patterns are matched
    - Require a dictionary extension
Syntactic Piece

- Our point
  - Propose a notion of *Syntactic piece*
  - Opinion extraction using Syntactic piece

- What is Syntactic Piece?
  - Minimum unit of syntactic structure
  - A pair consisting of a modifier and modifiee
  - This pair is expressed as follows
    \[ \text{Syntactic piece} : \text{modifier} \rightarrow \text{modifiee} \]
Bag of Words

Tree Structure

Word N-grams

Syntactic Pieces
Characteristics

- Very simple
  - It is easy to use, just like n-gram
- It has syntactic structure
  - It contains more information than n-gram
- Similar to phrasal idiom
  - It can deal with a chunk of meaning
- No need to switch domains
  - Existing works usually change dictionary to each domain
Method

1. Syntactic Piece Extraction
2. Calculate Pieces Score and Make Seed Dictionary
3. Dictionary Generalization
4. Sentence Classification
5. Dictionary Extension
Japanese sentence: シャープのケータイは画質がとてもいいです
(Cellular phone by SHARP picture quality is very good)
Calculate Pieces Score and Make Seed Dictionary

training corpus

positive

negative

extract

Syntactic Pieces

\{ phrase A → phrase B : positive \\
phrase C → phrase D : positive \\
phrase A → phrase B : negative \\
\ldots \\
\ldots \\
\ldots \}

calculate score

seed dictionary
Semantic Orientation Score

- Calculate piece score

\[
\text{score}(piece_i) = \frac{P(piece_i) - N(piece_i)}{P(piece_i) + N(piece_i)} \quad (-1 \leq \text{score}(piece_i) \leq 1)
\]

\[
\begin{cases}
\text{score}(piece_i) > 0 \rightarrow \text{positive phrase} \\
\text{score}(piece_i) < 0 \rightarrow \text{negative phrase}
\end{cases}
\]

\text{piece}_i \text{ is a syntactic piece.}
\text{score}(piece_i) \text{ is sentiment orientation score of } piece_i.
\text{P}(piece_i) \text{ is probability of } piece_i \text{ appeared in positive opinions.}
\text{N}(piece_i) \text{ is probability of } piece_i \text{ appeared in negative opinions.}
Dictionary Generalization

- Seed dictionary only use
  - Correspondences of the input piece
- Increase number of entries
  - Semantic orientation of a word may change with domain in many cases
  - However, some words always show only p or n
  - Extract modifier(modiffee) that always show only p or n
Dictionary Generalization

**Seed Dictionary**

- phrase B → phrase A
- phrase C → phrase A
- phrase D → phrase A
- phrase E → phrase A
- phrase G → phrase A
- phrase H → phrase A

**Generalized Dictionary**

- unknown → phrase A
- unknown → phrase A
- phrase D → phrase A
- phrase E → phrase A
- phrase G → phrase A
- phrase H → phrase A

**generalize**
Sentence Classification

Input sentence: (digital camera's design is good)

Extract:

Syntactic Piece: デジカメ-の → デザイン (digital camera's design)
Syntactic Piece: デザイン-が → 良い (design is good)

Compare:

any phrase → 良い (~ is good) : positive

Result:

Input sentence is positive opinion!
Sentence Classification

- Extract pieces in the input sentence
  - Only if dictionary dictionary have them
- Calculate sentence score

\[
sentence\text{ }score(S) = \sum_{piece_i \subset S} \text{score}(piece_i)
\]

\[
\begin{cases} 
\text{sentence score}(S) > 0 \rightarrow \text{positive opinion} \\
\text{sentence score}(S) < 0 \rightarrow \text{negative opinion} \\
\text{otherwise} \rightarrow \text{not opinion}
\end{cases}
\]

*piece_i is a syntactic piece in a sentence S.*

*sentence score(S) is its sentence score.*
Dictionary Extension

- Seed dictionary
  - Size of seed dictionary is small
    - Small dictionary gives low recall
    - If there is larger training corpus, size of seed dictionary can be large too.
    - But, not easy to increase training corpus by hand

- To improve recall
  - make a training corpus tagged p/n automatically
Dictionary Extension

- Seed dictionary
- Generalized dictionary
- General corpus
- Training
  - Not opinion
  - Don't use
- Extract
- Syntactic Pieces
  - Phrase A → Phrase B: positive
  - Phrase A → Phrase B: negative
- Calculate score
- New training corpus
- Extended dictionary
Experiment

- **Training corpus**
  - Weblogs (always tagged positive or negative)
  - 13 domains and 5,608 sentences

- **General corpus**
  - Weblogs (not tagged)
  - Million sentences

- **Evaluation**
  - 13-fold cross validation to each domains
Results and Discussion

- Result for sentence classification

<table>
<thead>
<tr>
<th>dictionary</th>
<th>precision</th>
<th>recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>seed only</td>
<td>0.85 (752/888)</td>
<td>0.13 (752/5608)</td>
</tr>
<tr>
<td>seed + generalization</td>
<td>0.86 (2423/2809)</td>
<td>0.43 (2423/5608)</td>
</tr>
<tr>
<td>extended seed</td>
<td>0.82 (1033/1257)</td>
<td>0.18 (1033/5608)</td>
</tr>
<tr>
<td>extension + generalization</td>
<td>0.91 (3046/3338)</td>
<td>0.54 (3046/5608)</td>
</tr>
</tbody>
</table>

Not decline precision even if dictionary is generalized

→ syntactic piece is effective unit
Results and Discussion

- Result for sentence classification

<table>
<thead>
<tr>
<th>dictionary</th>
<th>precision</th>
<th>recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>seed only</td>
<td>0.85 (752/888)</td>
<td>0.13 (752/5608)</td>
</tr>
<tr>
<td>seed + generalization</td>
<td>0.86 (2423/2809)</td>
<td>0.43 (2423/5608)</td>
</tr>
<tr>
<td>extended seed</td>
<td>0.82 (1033/1257)</td>
<td>0.18 (1033/5608)</td>
</tr>
<tr>
<td>extension + generalization</td>
<td>0.91 (3046/3338)</td>
<td>0.54 (3046/5608)</td>
</tr>
</tbody>
</table>

Low recall

Using larger general corpus → the recall can improve
# Result by each domains

<table>
<thead>
<tr>
<th>domain</th>
<th>precision</th>
<th>recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>digital camera</td>
<td>0.84 (408/484)</td>
<td>0.53 (408/771)</td>
</tr>
<tr>
<td>PC</td>
<td>0.90 (109/121)</td>
<td>0.51 (109/212)</td>
</tr>
<tr>
<td>soft drink</td>
<td>0.92 (406/441)</td>
<td>0.63 (406/649)</td>
</tr>
<tr>
<td>services</td>
<td>0.88 (206/233)</td>
<td>0.45 (206/456)</td>
</tr>
<tr>
<td>MP3 player</td>
<td>0.91 (317/350)</td>
<td>0.53 (317/595)</td>
</tr>
<tr>
<td>printer</td>
<td>0.91 (117/129)</td>
<td></td>
</tr>
<tr>
<td>cellular phone</td>
<td>0.96 (130/136)</td>
<td></td>
</tr>
<tr>
<td>designer goods</td>
<td>0.95 (156/164)</td>
<td></td>
</tr>
<tr>
<td>shampoo</td>
<td>0.91 (326/358)</td>
<td>0.56 (326/651)</td>
</tr>
<tr>
<td>beer</td>
<td>0.96 (544/567)</td>
<td>0.60 (544/909)</td>
</tr>
<tr>
<td>video game</td>
<td>0.89 (59/66)</td>
<td>0.52 (59/113)</td>
</tr>
<tr>
<td>cosmetics</td>
<td>1.00 (37/37)</td>
<td>0.66 (37/56)</td>
</tr>
<tr>
<td>sweets</td>
<td>0.92 (231/252)</td>
<td>0.55 (231/420)</td>
</tr>
</tbody>
</table>

High precision is obtained regardless of domains → no need to switch domains
## Results and Discussion

- **Result by each pattern**

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case frame</td>
<td>0.82 (417/506)</td>
<td>0.07 (417/5608)</td>
</tr>
<tr>
<td>Adverbial modification</td>
<td>0.85 (290/340)</td>
<td>0.05 (290/5608)</td>
</tr>
<tr>
<td>Verbal modification</td>
<td>0.88 (59/67)</td>
<td>0.01 (59/5608)</td>
</tr>
<tr>
<td>Adjectival modification</td>
<td>0.85 (69/81)</td>
<td>0.01 (69/5608)</td>
</tr>
<tr>
<td>Prefix</td>
<td>0.67 (16/24)</td>
<td>0.00 (16/5608)</td>
</tr>
</tbody>
</table>

**Adverbial modification pattern** is important in opinion extraction.
Conclusion

- Syntactic Piece is proposed
  - Minimum unit of syntactic structure
  - Easy to use, like n-gram
  - No need to switch domains

- Opinion extraction
  - Sentence classification using syntactic piece
  - Precision 91%, Recall 54%
Thank you
**Method**

**training**

- **training corpus**
  - classify sentence using dictionary
  - compare
- **dictionary**
  - extract pieces and calculate score
  - Syntactic Pieces: sentiment orientation
    - 画質-が ⇒ 良い (picture quality is good): positive
    - 値段-が ⇒ 高い (expensive): negative
    - 香り-が ⇒ 良い (good smell): positive
    - etc.
- count frequency and generalization
  - 良い (good): P:5, N:0
  - 悪い (bad): P:0, N:3
  - etc.

**test**

**Input sentence**

デジカメ-の デザイン-が 良い。
(digital camera's design is good)

**Syntactic Piece**

- デジカメ-の ⇒ デザイン (digital camera's design)
- デザイン-が ⇒ 良い (design is good)
- any phrase ⇒ 良い (good): positive

**Syntactic Piece**

- デザイン-が ⇒ 良い (design is good): positive

**Input sentence is positive opinion!**
Syntactic Piece

Japanese patterns of the piece

- Continuous modification
  - Case frame: noun(-particle) → predicate
    - 画面-が→きれい (clear screen)
  - Adverbial modification: adverb → predicate
    - とても→おいしい (delicious)

- Adnominal modification
  - Noun modification: noun(-no) → noun
    - キャノン-の→カメラ (canon's camera)
Syntactic Piece

- **Japanese patterns of the piece**
  - **Adnominal modification**
    - **Verbal modification**: verb → noun
      - くつろげる→店 (comfortable shop)
    - **Adjectival modification**: adjective → noun
      - おいしい→ケーキ (delicious cake)
  - **Compound noun**: noun-noun
    - 携帯-電話 (cellular phone)
  - **Prefix**: prefix-noun
    - 高-画質 (high picture quality)
## Example of positive pieces

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Syntactic Piece</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Case Frame</strong></td>
<td>コンテンツ-が⇒充実 (contents is enriched)</td>
</tr>
<tr>
<td></td>
<td>好感-を⇒持てる (favorable impression)</td>
</tr>
<tr>
<td></td>
<td>デザイン⇒が⇒かわいい (design is cute)</td>
</tr>
<tr>
<td></td>
<td>動作⇒が⇒速い (response is quick)</td>
</tr>
<tr>
<td></td>
<td>心地⇒良い (feel good)</td>
</tr>
<tr>
<td><strong>Verbal Modification</strong></td>
<td>暖まる⇒エピソード (heart warming episode)</td>
</tr>
<tr>
<td></td>
<td>楽しむ⇒方法 (way to enjoy)</td>
</tr>
<tr>
<td><strong>Adverbial Modification</strong></td>
<td>とっても⇒きれい (very beautiful)</td>
</tr>
<tr>
<td></td>
<td>かなり⇒コンパクト (very compact)</td>
</tr>
<tr>
<td><strong>Adjectival Modification</strong></td>
<td>いい⇒香り (good smell)</td>
</tr>
<tr>
<td></td>
<td>高い⇒品質 (high quality)</td>
</tr>
<tr>
<td></td>
<td>すごい⇒お洒落 (very stylish)</td>
</tr>
<tr>
<td><strong>Prefix</strong></td>
<td>新⇒商品 (new product)</td>
</tr>
<tr>
<td></td>
<td>省⇒スペース (small space)</td>
</tr>
</tbody>
</table>
### Example of negative pieces

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Syntactic piece</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case frame</td>
<td>画質が⇒良いない (picture quality is not good)</td>
</tr>
<tr>
<td></td>
<td>使い勝手が⇒悪い (usability is bad)</td>
</tr>
<tr>
<td></td>
<td>消耗が⇒激しい (very waste)</td>
</tr>
<tr>
<td></td>
<td>サイズが⇒小さい (size is small)</td>
</tr>
<tr>
<td></td>
<td>気持ち⇒悪い (feel sick)</td>
</tr>
<tr>
<td>Verbal modification</td>
<td>違う⇒商品 (different item)</td>
</tr>
<tr>
<td>Adverbial modification</td>
<td>すぐ⇒壊れる (break at once)</td>
</tr>
<tr>
<td></td>
<td>かなり⇒高額 (very extensive)</td>
</tr>
<tr>
<td>Adjectival modification</td>
<td>ぬるい⇒ビール (lukewarm beer)</td>
</tr>
<tr>
<td></td>
<td>物足りない⇒感じ (not good enough)</td>
</tr>
<tr>
<td>Prefix</td>
<td>異⇒音 (noise)</td>
</tr>
<tr>
<td></td>
<td>再⇒起動 (reboot)</td>
</tr>
</tbody>
</table>
Example of generalized dictionary

<table>
<thead>
<tr>
<th>semantic orientation</th>
<th>syntactic piece</th>
</tr>
</thead>
</table>
| positive             | any phrase ⇒ キレイ (beautiful)  
any phrase ⇒ 使いやすい (easy to use)  
any phrase ⇒ 美味しい (good taste)  
飲みやすい (easy to drink) ⇒ any phrase |
| negative             | any phrase ⇒ 良いな (no good)  
any phrase ⇒ 使いにくい (hard to use)  
any phrase ⇒ まずい (bad taste)  
いまひとつ (unattractive) ⇒ any phrase |
# Experiment

## Number of sentence

<table>
<thead>
<tr>
<th>domain</th>
<th>positive</th>
<th>negative</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>digital camera</td>
<td>533</td>
<td>238</td>
<td>771</td>
</tr>
<tr>
<td>PC</td>
<td>112</td>
<td>100</td>
<td>212</td>
</tr>
<tr>
<td>soft drink</td>
<td>559</td>
<td>90</td>
<td>649</td>
</tr>
<tr>
<td>services</td>
<td>185</td>
<td>271</td>
<td>456</td>
</tr>
<tr>
<td>MP3 player</td>
<td>364</td>
<td>231</td>
<td>595</td>
</tr>
<tr>
<td>printer</td>
<td>103</td>
<td>177</td>
<td>280</td>
</tr>
<tr>
<td>cellular phone</td>
<td>156</td>
<td>73</td>
<td>229</td>
</tr>
<tr>
<td>designer goods</td>
<td>221</td>
<td>46</td>
<td>267</td>
</tr>
<tr>
<td>shampoo</td>
<td>478</td>
<td>173</td>
<td>651</td>
</tr>
<tr>
<td>beer</td>
<td>748</td>
<td>161</td>
<td>909</td>
</tr>
<tr>
<td>video game</td>
<td>61</td>
<td>52</td>
<td>113</td>
</tr>
<tr>
<td>cosmetics</td>
<td>44</td>
<td>12</td>
<td>56</td>
</tr>
<tr>
<td>sweets</td>
<td>322</td>
<td>98</td>
<td>420</td>
</tr>
</tbody>
</table>
Abstract

- **Purpose**
  - Opinion Extraction from given document
  - Positive / Negative classification
- **My Point**
  - Propose a notion of *Syntactic piece*
  - Opinion extraction using Syntactic piece
Method

- Semantic orientation score
  - Sentiment orientation
    - Positive phrase appear in positive opinion
    - The same can be said syntactic piece
Dictionary Generalization

Example of generalization

Positive

- 画質-が→良い (picture quality is good)
- 味-が→良い (taste good)
- 画面-が→大きい (screen is big)

Negative

- 騒音-が→大きい (noise is big)
- デザイン-が→悪い (design is bad)
- 印象-が→悪い (impression is bad)

Any phrase→良い (good) : tagged positive

Any phrase→悪い (bad) : tagged negative
Dictionary Extension

- Make a new training corpus
  - Use seed and generalized dictionary
  - Classify general corpus (positive/negative/other)
- Extended dictionary
  - Extract pieces in new training corpus
  - Calculate piece score
  - Add this pieces into dictionary
Calculate Pieces Score and Make Seed Dictionary

Training corpus:
- Extract pieces and calculate score
- Classify sentence using dictionary
- Compare

Dictionary:
- Syntactic Pieces: sentiment orientation
  - 画質-が → 良い (picture quality is good): positive
  - 値段-が → 高い (expensive): negative
  - 香り-が → 良い (good smell): positive
  - etc.

General corpus:
- Syntactic Pieces: sentiment orientation
  - 良い (good): P:5, N:0
  - 悪い (bad): P:0, N:3
  - etc.
Dictionary Generalization

training corpus

positive

negative

dictinary

general corpus

extract pieces and calculate score

classify sentence using dictionary

compare

training

Syntactic Pieces: sentiment orientation

良い (good): positive
悪い (bad): negative

等 (etc.)

良い (good) P:5, N:0
悪い (bad) P:0, N:3

Syntactic Pieces: sentiment orientation

picture quality is good
expensive
good smell

equal or similar phrases
Dictionary Extension

positive

negative

training corpus

classify sentence using dictionary

count frequency and generalization

extract pieces and calculate score

Syntactic Pieces: sentiment orientation

良い (good) → 良い (positive)
悪い (bad) → 悪い (negative)

textbook

dictionary

良い (good) P:5, N:0
悪い (bad) P:0, N:3

equal

any phrase → 良い (good): positive
any phrase → 悪い (bad): negative

equal

general corpus

Syntactic Pieces: sentiment orientation

画質が → 良い (picture quality is good): positive
値段が → 高い (expensive): negative
香りが → 良い (good smell): positive

equal