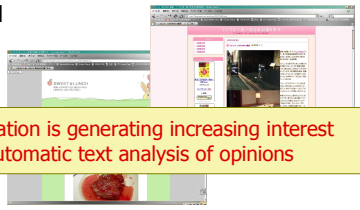
Building an opinion-tagged corpus for opinion extraction

Nozomi Kobayashi
Computational linguistics lab.
Nara Institute of Science and Technology

Background (1/2)

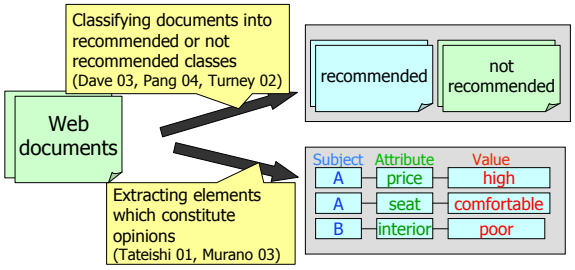
- There are an increasing number of Web documents that include human opinions
  - These information are scattered on the web
  - Acquiring users' opinions on products or services is helpful



This situation is generating increasing interest in automatic text analysis of opinions

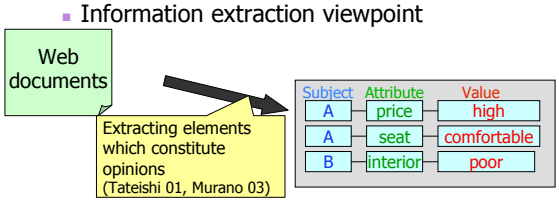
Background (2/2)

- The task of opinion acquisition:



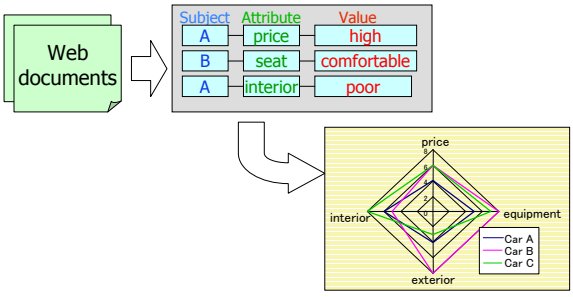
Aim

- To extract opinions that describe evaluation of particular products together with evidence
  - Information extraction viewpoint



One example of Application

- Generation of radar charts



Task definition (to date)

- Opinions are often expressed in the form of an attribute-value pair

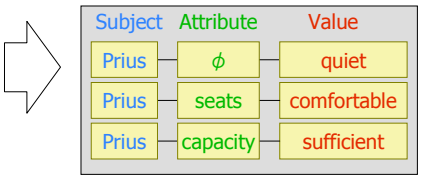
Attribute: one aspect of a product (subject)
(e.g. engine, design, price,…)
Value: specific expression that quantifies or qualifies the aspect
(e.g. high, good, beautiful, …)

The task: extracting ⟨subject, attribute, value⟩

Task definition (to date)

- Examples of ⟨subject, attribute, value⟩ extraction

I bought the Prius last month.
The car is very quiet and the seats are comfortable.
The capacity of the trunk may be sufficient.



7

Extending the current framework

- Large variety of ways of expressing opinions
  - Hierarchical relations in "Attributes"

    the capacity of the trunk
  - Comparisons and conditions

    The car gets good mileage at high speed.
  - Reasons

    This car is practical since we can change the seat arrangement

8

Outline

- Background and aim
- Task definition (to date)
- Building an opinion-tagged corpus
  - Procedure for annotation
  - Statistics of opinion-tagged corpus
  - Inter-annotator agreement
- Conclusion and future direction

10

Opinion-tagged corpus

- To investigate how many opinions have hierarchical relations
  - Annotating relation information
    - attribute-value relation

      The sound of the engine is excellent
    - attribute-attribute relation

      The sound of the engine is excellent

11

Procedure for annotation

- Identify candidate values using a dictionary

  Mazda axela, I like the design. But, the capacity of the trunk is insufficient. Hmm…If only the capacity were larger..
- Identify the relation between a value and the attribute

  Mazda axela, I like the design. But, the capacity of the trunk is insufficient. Hmm…If only the capacity were larger..
- Identify the hierarchical relation between attributes

  Mazda axela, I like the design. But, the capacity of the trunk is insufficient. Hmm…If only the capacity were larger..

Value dictionary

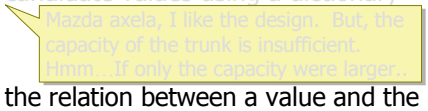
- Collecting the value expressions:
  - Semi-automatic method (2,638 expressions)
  - Existing resources
    - Japanese thesauruses (2,853 expressions)
    - Iwanami Japanese dictionary (129 expressions)

Total number of expressions : 5,620

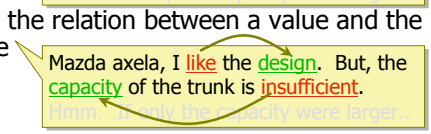
14

Procedure for annotation

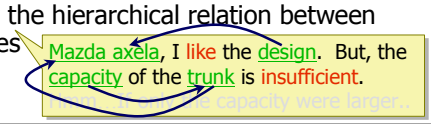
- Identify candidate values using a dictionary



- Identify the relation between a value and the attribute



- Identify the hierarchical relation between attributes



Building an opinion-tagged corpus

- Japanese weblog articles
  - 4 domain
    - restaurant, automobile, cellphone, and videogame
- We hired a person as the annotator
  - Annotating the relations using annotation tool

16

Statistics of opinion-tagged corpus

	restaurant	automobile	cellphone	videogame
# of articles	1,445	564	494	361
# of sentences	25,500	14,593	12,326	6,823
# of A-V pairs	4,504	1,017	1,144	551
# of A-A relations	2,054	280	304	221

A-V: attribute-value
A-A: attribute-attribute

17

Inter-annotator agreement

- 116 blog articles (2,805 sentences)
  - restaurant domain
  - candidate values are previously identified
- To evaluate the inter-annotator agreement
  - Use kappa statistics

18

Kappa statistics

- A measure of agreement that exists beyond the amount expected by chance
- Values of kappa: from 0 to 1
  - 1: perfect agreement
  - 0: chance alone

Kappa = (agreement — chance) / chance

19

Inter-annotator agreement (value)

- attribute-value relation

	value	not value
annotator1	374	1365
annotator2	337	1402
agree		1580 / 1756

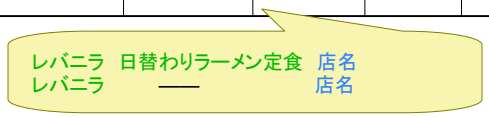
- Kappa statistics = 0.70

20

Inter-annotator agreement

- attribute-attribute relation

	completely matched	partially matched	others	total
annotator1	157	104	132	393
annotator2	157	104	109	370



21

Conclusion

- For extending the subject-attribute-value extraction
  - Identification of the relations between attributes
- For this identification
  - We build an opinion-tagged corpus
  - In this annotation scheme
    - Inter-annotator agreement is 0.7 (kappa)

22

Future direction

- We consider the task of extracting attribute-attribute relations
  - Use statistical information
    - "termhood" of the candidates
    - the degree of co-occurrence between candidate attributes

23