Opinion Mining and Summarization

Bing Liu
University Of Illinois at Chicago
liub@cs.uic.edu
http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

Introduction

- Two main types of textual information.
  - Facts and Opinions
- Most current text information processing methods (e.g., web search, text mining) work with factual information.
- Opinion mining or sentiment analysis
  - computational study of opinions, sentiments and emotions expressed in text.
- Why opinion mining now? Mainly because of the Web; huge volumes of opinionated text.

Introduction – user-generated media

- Importance of opinions:
  - Opinions are so important that whenever we need to make a decision, we want to hear others' opinions.
  - In the past,
    - Individuals: opinions from friends and family
    - businesses: surveys, focus groups, consultants …
- Word-of-mouth on the Web
  - User-generated media: One can express opinions on anything in reviews, forums, discussion groups, blogs ...
  - Opinions of global scale: No longer limited to:
    - Individuals: one's circle of friends
    - Businesses: Small scale surveys, tiny focus groups, etc.

A Fascinating Problem!

- Intellectually challenging & major applications.
  - A very popular research topic in recent years in NLP and Web data mining.
  - 20-60 companies in USA alone
- It touches everything aspect of NLP and yet is restricted and confined.
  - Little research in NLP/Linguistics in the past.
- Potentially a major technology from NLP.
  - But it is not easy!

Roadmap

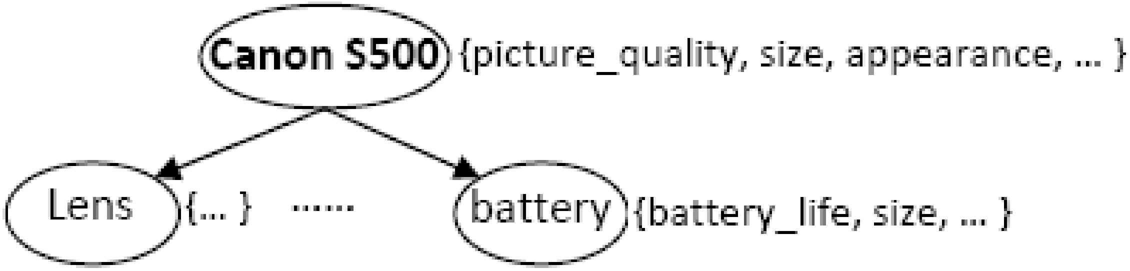
- **Opinion mining – problem definition**
- Document level sentiment classification
- Sentence level sentiment classification
- Opinion lexicon generation
- Feature-based opinion mining
- Opinion mining of comparative sentences
- Opinion spam detection
- Summary

An Example Review

- *"I bought an iPhone a few days ago. It was such a nice phone. The touch screen was really cool. The voice quality was clear too. Although the battery life was not long, that is ok for me. However, my mother was mad with me as I did not tell her before I bought the phone. She also thought the phone was too expensive, and wanted me to return it to the shop. …"*
- What do we see?
  - **Opinions, targets of opinions, and opinion holders**

Target Object (Liu, Web Data Mining book, 2006)

- **Definition** (**object**): An *object o* is a product, person, event, organization, or topic. *o* is represented as
  - a hierarchy of components, sub-components, and so on.
  - Each node represents a component and is associated with a set of attributes of the component.


Canon S500
{picture_quality, size, appearance, ... }
Lens
{... }
......
battery
{battery_life, size, ... }

- An opinion can be expressed on any node or attribute of the node.
- To simplify our discussion, we use the term ***features*** to represent both components and attributes.

What is an Opinion? (Liu, Ch. in NLP handbook)

- An *opinion* is a quintuple

  (o_j, f_{jk}, so_{ijkl} , h_i, t_l),

  where

  - o_j is a target object.
  - f_{jk} is a feature of the object o_j.
  - so_{ijkl} is the sentiment value of the opinion of the opinion holder h_i on feature f_{jk} of object o_j at time t_l. so_{ijkl} is +ve, -ve, or neu, or a more granular rating.
  - h_i is an opinion holder.
  - t_l is the time when the opinion is expressed.

Objective – structure the unstructured

- Objective: Given an opinionated document,
  - Discover all quintuples (o_j, f_{jk}, so_{ijkl} , h_i, t_l),
    - i.e., mine the five corresponding pieces of information in each quintuple, and
  - Or, solve some simpler problems
- With the quintuples,
  - Unstructured Text → Structured Data
    - Traditional data and visualization tools can be used to slice, dice and visualize the results in all kinds of ways
    - Enable qualitative and quantitative analysis.

Feature-Based Opinion Summary

(Hu & Liu, KDD 2004)

"I bought an iPhone a few days ago. It was such a nice phone. The touch screen was really cool. The voice quality was clear too. Although the battery life was not long, that is ok for me. However, my mother was mad with me as I did not tell her before I bought the phone. She also thought the phone was too expensive, and wanted me to return it to the shop. …"

….

Feature Based Summary:

Feature1: **Touch screen**

Positive: 212

- *The touch screen was really cool.*
- *The touch screen was so easy to use and can do amazing things.*

…

Negative: 6

- The screen is easily scratched.
- I have a lot of difficulty in removing finger marks from the touch screen.

…

Feature2: **battery life**

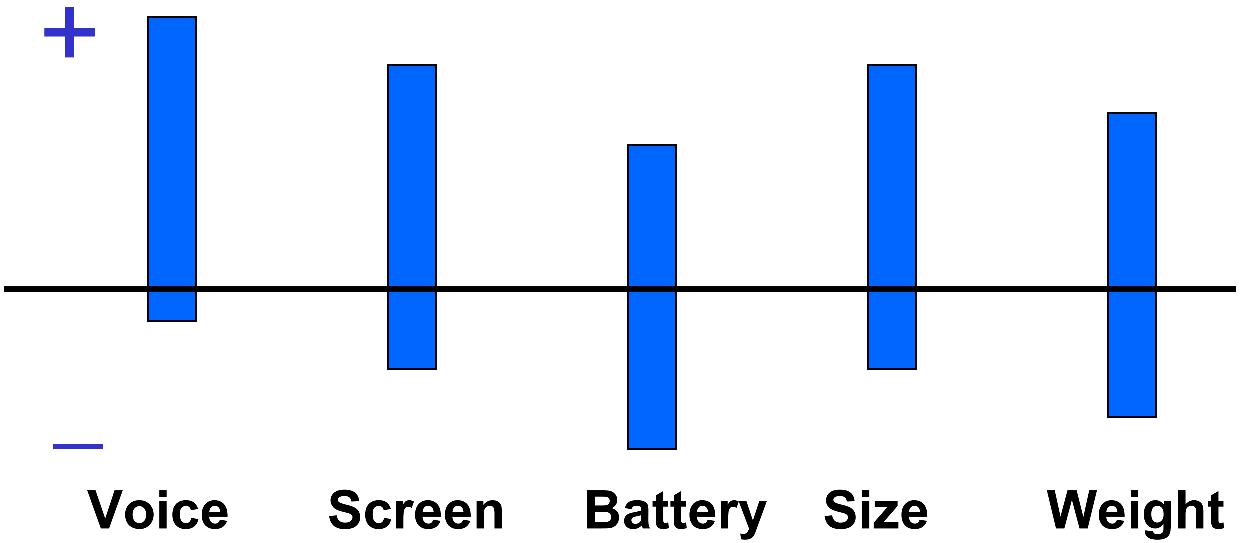
…

Note: We omit opinion holders

Visual Comparison (Liu et al. WWW-2005)

- Summary of reviews of

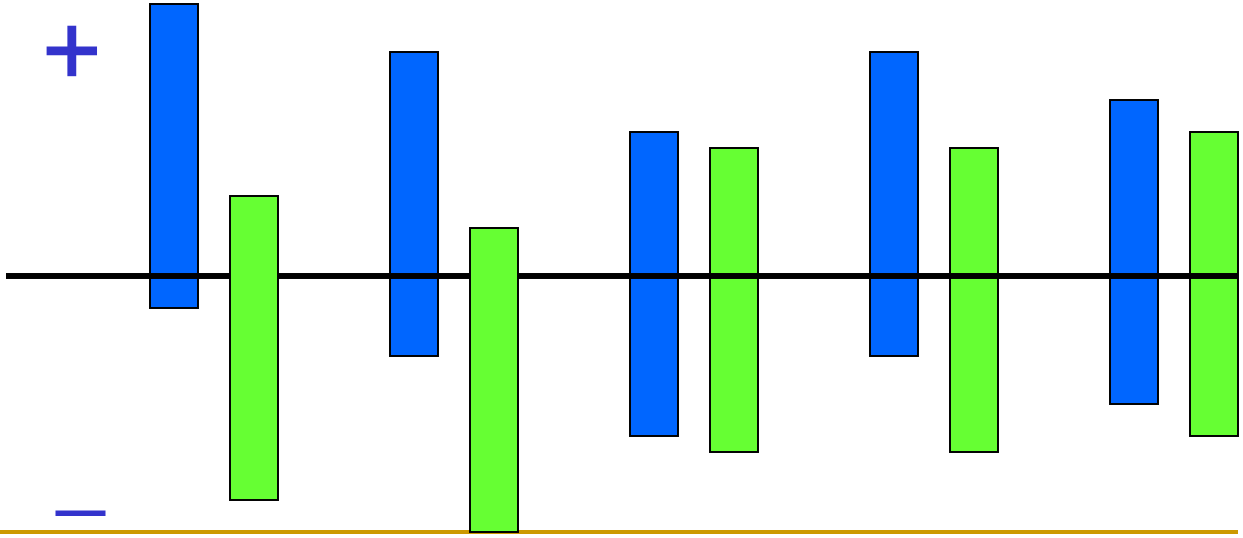
🟦 Cell Phone 1



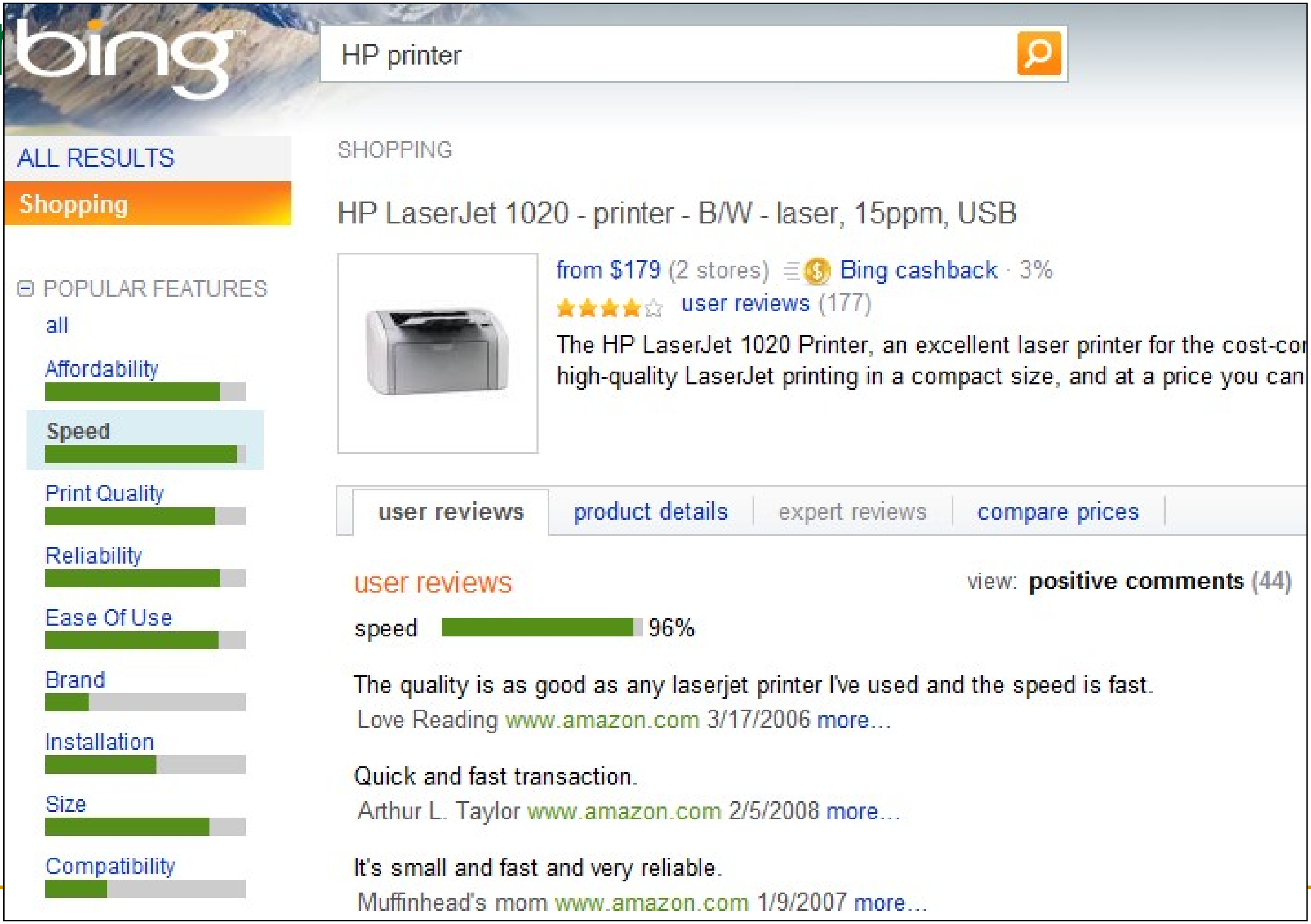
- Comparison of reviews of

🟦 Cell Phone 1

🟩 Cell Phone 2



Feat.-based opinion summary i



bing

HP printer

ALL RESULTS

Shopping

⊟ POPULAR FEATURES

all

Affordability

Speed

Print Quality

Reliability

Ease Of Use

Brand

Installation

Size

Compatibility

SHOPPING

HP LaserJet 1020 - printer - B/W - laser, 15ppm, USB

from $179 (2 stores) Bing cashback · 3%

★★★★☆ user reviews (177)

The HP LaserJet 1020 Printer, an excellent laser printer for the cost-co
high-quality LaserJet printing in a compact size, and at a price you can

user reviews | product details | expert reviews | compare prices

user reviews

view: **positive comments** (44)

speed 96%

The quality is as good as any laserjet printer I've used and the speed is fast.
Love Reading www.amazon.com 3/17/2006 more...

Quick and fast transaction.
Arthur L. Taylor www.amazon.com 2/5/2008 more...

It's small and fast and very reliable.
Muffinhead's mom www.amazon.com 1/9/2007 more...

Opinion Mining is Hard!

- *"This past Saturday, I bought a Nokia phone and my girlfriend bought a Motorola phone with Bluetooth. We called each other when we got home. The voice on my phone was not so clear, worse than my previous phone. The battery life was long. My girlfriend was quite happy with her phone. I wanted a phone with good sound quality. So my purchase was a real disappointment. I returned the phone yesterday."*

It is not Just ONE Problem

- $(o_j, f_{jk}, so_{ijkl}, h_i, t_l)$,
  - o_j - a target object: Named Entity Extraction (more)
  - f_{jk} - a feature of o_j: Information Extraction
  - so_{ijkl} is sentiment: Sentiment determination
  - h_i is an opinion holder: Information/Data Extraction
  - t_l is the time: Data Extraction
- Co-reference resolution
- Synonym match (voice = sound quality) …
- None of them is a solved problem!

Roadmap

- Opinion mining – problem definition
- **Document level sentiment classification**
- Sentence level sentiment classification
- Opinion lexicon generation
- Feature-based opinion mining
- Opinion mining of comparative sentences
- Opinion spam detection
- Summary

Sentiment Classification: doc-level (Pang and Lee, Survey, 2008)

- Classify a document (e.g., a review) based on the overall sentiment expressed by opinion holder
  - Classes: Positive, or negative
- Assumption: each document focuses on a single object and contains opinions from a single op. holder.
- *E.g., thumbs-up or thumbs-down?*
  - *"I bought an iPhone a few days ago. It was such a nice phone. The touch screen was really cool. The voice quality was clear too. Although the battery life was not long, that is ok for me. However, my mother was mad with me as I did not tell her before I bought the phone. She also thought the phone was too expensive, and wanted me to return it to the shop. …"*

Sentiment classification

- Classify a document (e.g., a product reviews) based on the overall sentiment expressed by opinion holder
  - Classes: Positive or negative
  - Since in our model an object *O* itself is also a feature, then sentiment classification essentially determines the opinion expressed on *O* in each document (e.g., review).
- Assumption: each document (or review) focuses on a single object and contains opinions from a single opinion holder.
  - Not always true, e.g., forum postings, and blogs

Unsupervised sentiment classification (Turney, ACL-02)

- Data: reviews from epinions.com on automobiles, banks, movies, and travel destinations.
- The approach: Three steps
- Step 1:
  - Part-of-speech tagging
  - Extracting two consecutive words (two-word phrases) from reviews if their tags conform to some given patterns, e.g., (1) JJ, (2) NN.

- Step 2: Estimate the semantic orientation (SO) of the extracted phrases
  - Use Pointwise mutual information

$$PMI(word_1, word_2) = \log_2\left(\frac{P(word_1 \wedge word_2)}{P(word_1)P(word_2)}\right)$$

  - *SO(phrase) = PMI(phrase, “excellent”)*
    - PMI(phrase, “poor”)
- Step 3: Compute the average SO of all phrases
  - classify the review as recommended if average SO is positive, not recommended otherwise.

Supervised sentiment classification

(Pang et al, EMNLP-02)

- This paper directly applied several machine learning techniques to classify movie reviews into positive and negative.
- Three classification techniques were tried:
  - Naïve Bayes
  - Maximum entropy
  - Support vector machine
- Pre-processing settings: negation tag, unigram (single words), bigram, POS tag, position.
- SVM: the best accuracy 83% (unigram)

Roadmap

- Opinion mining – problem definition
- Document level sentiment classification
- **Sentence level sentiment classification**
- Opinion lexicon generation
- Feature-based opinion mining
- Opinion mining of comparative sentences
- Opinion spam detection
- Summary

Subjectivity Analysis: sent.-level (Wiebe et al 2004)

- Sentence-level sentiment analysis has two tasks:
  - Subjectivity classification: Subjective or objective.
    - Objective: e.g., *I bought an iPhone a few days ago.*
    - Subjective: e.g., *It is such a nice phone.*
  - Sentiment classification: For subjective sentences or clauses, classify positive or negative.
    - Positive: *It is such a nice phone.*
- But (Liu, a Ch in NLP handbook)
  - subjective sentences ≠ +ve or –ve opinions
    - E.g., *I think he came yesterday.*
  - Objective sentence ≠ no opinion
    - Imply –ve opinion: *The phone broke in two days*

Sentence-level sentiment analysis

- Document-level sentiment classification is too coarse for many applications.
- We move to the sentence level.
- Much of the work on sentence level sentiment analysis focuses on identifying subjective sentences in news articles.
  - Classification: objective and subjective.
  - All techniques use some forms of machine learning.
  - E.g., using a naïve Bayesian classifier with a set of data features/attributes extracted from training sentences (Wiebe et al. ACL-99).

Using learnt patterns (Rilloff and Wiebe, EMNLP-03)

- A bootstrapping approach.
  - A high precision classifier is first used to automatically identify some subjective and objective sentences.
    - Two high precision (but low recall) classifiers are used,
      - a high precision subjective classifier
      - A high precision objective classifier
      - Based on manually collected lexical items, single words and n-grams, which are good subjective clues.
  - A set of patterns are then learned from these identified subjective and objective sentences.
    - Syntactic templates are provided to restrict the kinds of patterns to be discovered, e.g., <subj> passive-verb.
  - The learned patterns are then used to extract more subject and objective sentences (the process can be repeated).

Roadmap

- Opinion mining – problem definition
- Document level sentiment classification
- Sentence level sentiment classification
- **Opinion lexicon generation**
- Feature-based opinion mining
- Opinion mining of comparative sentences
- Opinion spam detection
- Summary

Opinion words and phrases

- Opinion words or phrases (*opinion lexicon*): they are crucial for opinion mining (obviously!):
  - Positive: beautiful, wonderful, good, amazing,
  - Negative: bad, poor, terrible, cost someone an arm and a leg.
- Three main ways to compile such a list:
  - Manual approach: not a bad idea, only an one-time effort
  - Corpus-based approaches
  - Dictionary-based approaches
- Important to note:
  - Some opinion words are context independent (e.g., good).
  - Some are context dependent (e.g., long, cheap).

Corpus-based approaches

- Use constraints (or conventions) on connectives to identify opinion words (Hazivassiloglou and McKeown, ACL-97; Kanayama and Nasukawa, EMNLP-06; Ding and Liu, 2007).
- Conjunction: conjoined adjectives usually have the same orientation (Hazivassiloglou and McKeown, ACL-97).
  - E.g., "This car is *beautiful* **and** *spacious*." (conjunction)
  - If we know beautiful is positive, spacious is ikely to be positive.
- AND, OR, BUT, EITHER-OR, and NEITHER-NOR have similar constraints.
- Learning and clustering
- Corpus: 21 million word 1987 Wall Street Journal corpus.

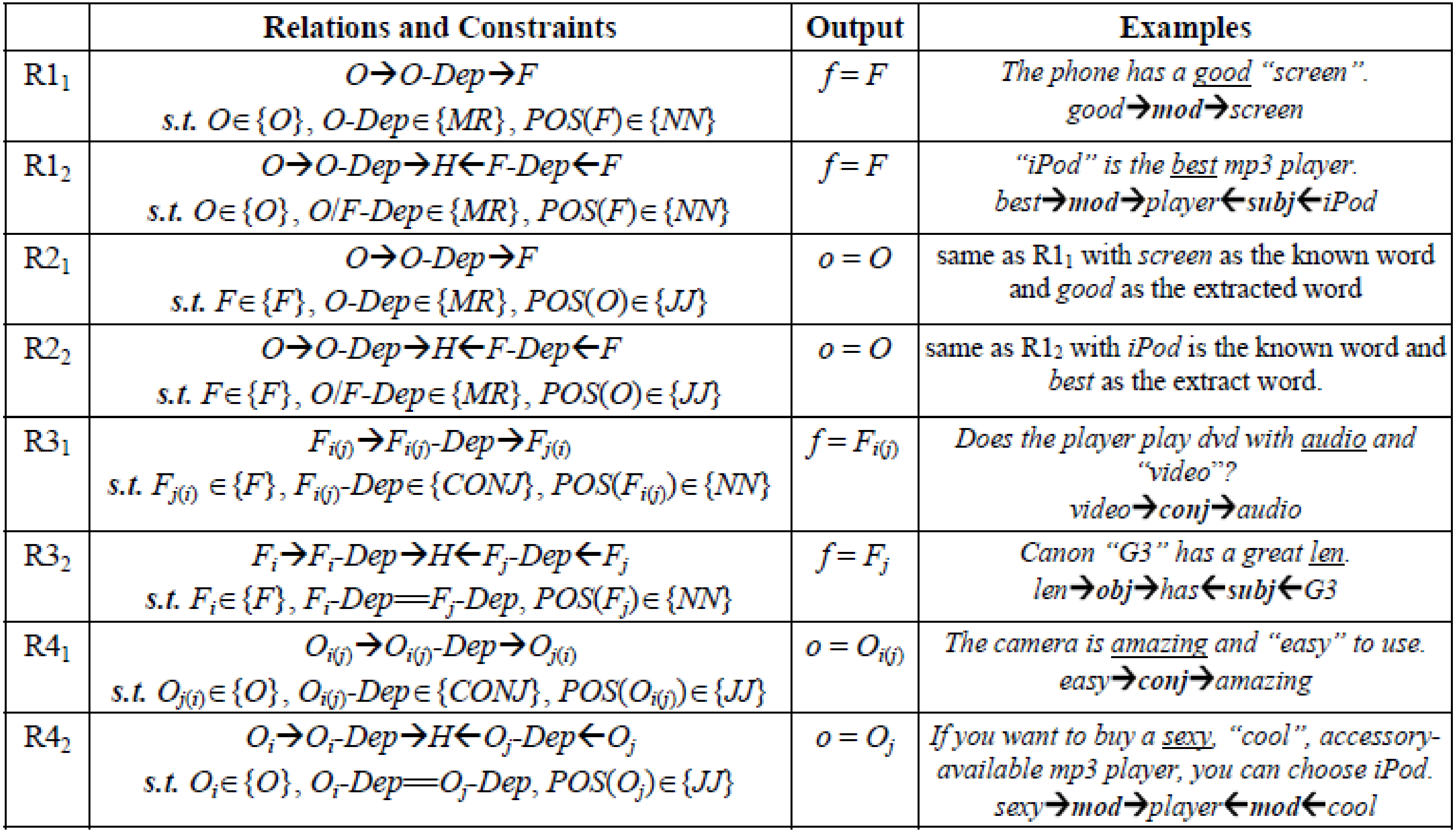
Corpus-based approaches (contd)

- (Kanayama and Nasukawa, EMNLP-06) takes a similar approach to (Hazivassiloglou and McKeown, ACL-97) but for Japanese words:
  - Instead of using learning, it uses two criteria to determine whether to add a word to positive or negative lexicon.
  - Have an initial seed lexicon of positive and negative words.
- (Ding and Liu, 2007) also exploits constraints on connectives, but with a key difference
  - It uses them to assign opinion orientations to product features (more on this later).
    - One word may indicate different opinions in the same domain.
      - "The battery life is *long*" (+) and "It takes a *long* time to focus" (-).
    - **Find domain opinion words is insufficient.**

Corpus-based approaches (contd)

- A double propagation method is proposed in [Qiu et al. IJCAI-2009]
- It exploits dependency relations of opinions and features to extract opinion words.
  - Opinions words modify object features, e.g.,
  - "This camera has *long* *battery life*"
- The algorithm essentially bootstraps using a set of seed opinion words
  - With the help of some dependency relations.

Rules from dependency grammar



	Relations and Constraints	Output	Examples
R1$_1$	$O \rightarrow O\text{-}Dep \rightarrow F$ s.t. $O \in \{O\}$, $O\text{-}Dep \in \{MR\}$, $POS(F) \in \{NN\}$	$f = F$	*The phone has a good "screen".* *good→mod→screen*
R1$_2$	$O \rightarrow O\text{-}Dep \rightarrow H \leftarrow F\text{-}Dep \leftarrow F$ s.t. $O \in \{O\}$, $O/F\text{-}Dep \in \{MR\}$, $POS(F) \in \{NN\}$	$f = F$	*"iPod" is the best mp3 player.* *best→mod→player←subj←iPod*
R2$_1$	$O \rightarrow O\text{-}Dep \rightarrow F$ s.t. $F \in \{F\}$, $O\text{-}Dep \in \{MR\}$, $POS(O) \in \{JJ\}$	$o = O$	same as R1$_1$ with *screen* as the known word and *good* as the extracted word
R2$_2$	$O \rightarrow O\text{-}Dep \rightarrow H \leftarrow F\text{-}Dep \leftarrow F$ s.t. $F \in \{F\}$, $O/F\text{-}Dep \in \{MR\}$, $POS(O) \in \{JJ\}$	$o = O$	same as R1$_2$ with *iPod* is the known word and *best* as the extract word.
R3$_1$	$F_{i(j)} \rightarrow F_{i(j)}\text{-}Dep \rightarrow F_{j(i)}$ s.t. $F_{j(i)} \in \{F\}$, $F_{i(j)}\text{-}Dep \in \{CONJ\}$, $POS(F_{i(j)}) \in \{NN\}$	$f = F_{i(j)}$	*Does the player play dvd with audio and "video"?* *video→conj→audio*
R3$_2$	$F_i \rightarrow F_i\text{-}Dep \rightarrow H \leftarrow F_j\text{-}Dep \leftarrow F_j$ s.t. $F_i \in \{F\}$, $F_i\text{-}Dep == F_j\text{-}Dep$, $POS(F_j) \in \{NN\}$	$f = F_j$	*Canon "G3" has a great len.* *len→obj→has←subj←G3*
R4$_1$	$O_{i(j)} \rightarrow O_{i(j)}\text{-}Dep \rightarrow O_{j(i)}$ s.t. $O_{j(i)} \in \{O\}$, $O_{i(j)}\text{-}Dep \in \{CONJ\}$, $POS(O_{i(j)}) \in \{JJ\}$	$o = O_{i(j)}$	*The camera is amazing and "easy" to use.* *easy→conj→amazing*
R4$_2$	$O_i \rightarrow O_i\text{-}Dep \rightarrow H \leftarrow O_j\text{-}Dep \leftarrow O_j$ s.t. $O_i \in \{O\}$, $O_i\text{-}Dep == O_j\text{-}Dep$, $POS(O_j) \in \{JJ\}$	$o = O_j$	*If you want to buy a sexy, "cool", accessory-available mp3 player, you can choose iPod.* *sexy→mod→player←mod←cool*

Dictionary-based approaches

- Typically use WordNet's synsets and hierarchies to acquire opinion words
  - Start with a small seed set of opinion words.
  - Use the set to search for synonyms and antonyms in WordNet (Hu and Liu, KDD-04; Kim and Hovy, COLING-04).
  - Manual inspection may be used afterward.
- Use additional information (e.g., glosses) from WordNet (Andreevskaia and Bergler, EACL-06) and learning (Esuti and Sebastiani, CIKM-05).
- Weakness of the approach: Do not find context dependent opinion words, e.g., small, long, fast.

Roadmap

- Opinion mining – problem definition
- Document level sentiment classification
- Sentence level sentiment classification
- Opinion lexicon generation
- **Feature-based opinion mining**
- Opinion mining of comparative sentences
- Opinion spam detection
- Summary

Feature-Based Sentiment Analysis

(Hu and Liu, KDD-2004)

- Sentiment classification at both document and sentence (or clause) levels are not enough,
  - they do not tell what people like and/or dislike
  - A positive opinion on an object does not mean that the opinion holder likes everything.
  - An negative opinion on an object does not mean …..
- Objective (recall): Discovering all quintuples

  $(o_j, f_{jk}, so_{ijkl}, h_i, t_l)$

- With all quintuples, all kinds of analyses become possible.

Feature-based opinion mining

- Recall: Mining all quintuples

  (H_i, O_j, f_{jk}, T_l, P_{ijkl}),

  where

  - H_i is an opinion holder,
  - O_j is an object,
  - f_{jk} is a feature of the object O_j,
  - T_l is the time when the opinion is expressed, and
  - P_{ijkl} is the orientation or polarity of the opinion of the opinion holder H_i on feature f_{jk} of object O_j at time T_l. P_{ijkl} is positive, negative or neutral.

The tasks

- Using product reviews as opinionated texts, we have three main tasks.

  Task 1: Extract object features that have been commented on in each review.

  Task 2: Determine whether the opinions on the features are positive, negative or neutral.

  Task 3: Group feature synonyms.

  - Produce a summary

Feature extraction

- Frequency-based approach (Hu and Liu, KDD-04):
- Find frequent features: those features that have been talked about by many reviewers.
- Use sequential pattern mining
- Why the frequency based approach?
  - Different reviewers tell different stories (irrelevant)
  - When product features are discussed, the words that they use converge.
  - They are main features.
- Sequential pattern mining finds frequent phrases.

Using part-of relationship and the Web

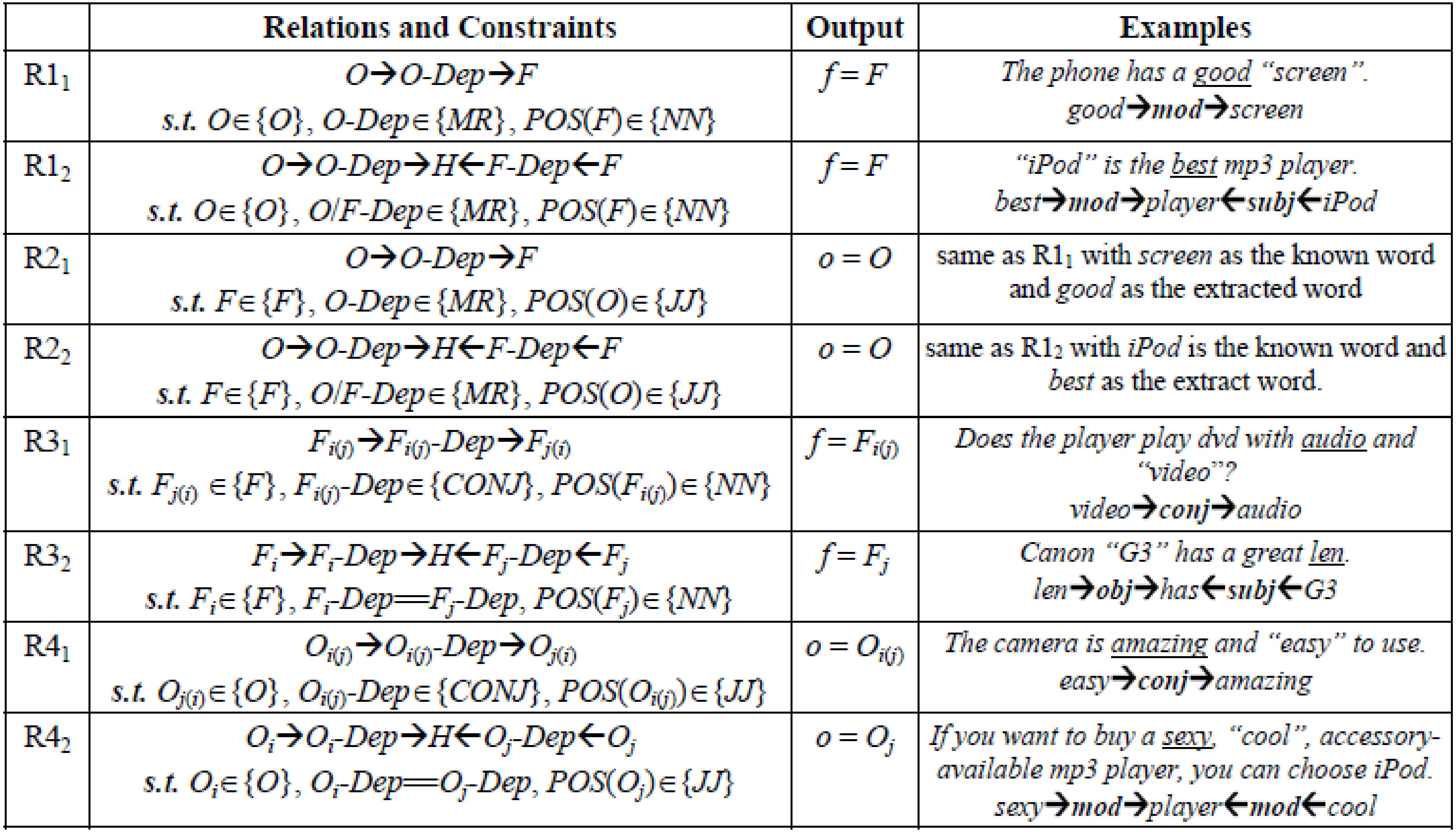
(Popescu and Etzioni, EMNLP-05)

- Improved (Hu and Liu, KDD-04) by removing those frequent noun phrases that may not be features: better precision (a small drop in recall).
- It identifies part-of relationship
  - Each noun phrase is given a PMI score between the phrase and **part discriminators** associated with the product class, e.g., a scanner class.
  - The part discriminators for the scanner class are, “of scanner”, “scanner has”, “scanner comes with”, etc, which are used to find components or parts of scanners by searching on the Web (Etzioni et al, WWW-04).

Using dependency relations
(Qui et al. IJCAI-2009)

- A same *double propagation* approach in (Qiu et al. IJCAI-2009) is applicable here.
- It exploits the dependency relations of opinions and features to extract features.
  - Opinions words modify object features, e.g.,
  - “This camera has *long* *battery life*”
- The algorithm bootstraps using a set of seed opinion words (no feature input).
  - To extract features (and also opinion words)

Rules from dependency grammar



	Relations and Constraints	Output	Examples
R1$_1$	$O \rightarrow O\text{-}Dep \rightarrow F$ s.t. $O \in \{O\}$, $O\text{-}Dep \in \{MR\}$, $POS(F) \in \{NN\}$	$f = F$	*The phone has a good "screen".* *good*→*mod*→*screen*
R1$_2$	$O \rightarrow O\text{-}Dep \rightarrow H \leftarrow F\text{-}Dep \leftarrow F$ s.t. $O \in \{O\}$, $O/F\text{-}Dep \in \{MR\}$, $POS(F) \in \{NN\}$	$f = F$	*"iPod" is the best mp3 player.* *best*→*mod*→*player*←*subj*←*iPod*
R2$_1$	$O \rightarrow O\text{-}Dep \rightarrow F$ s.t. $F \in \{F\}$, $O\text{-}Dep \in \{MR\}$, $POS(O) \in \{JJ\}$	$o = O$	same as R1$_1$ with *screen* as the known word and *good* as the extracted word
R2$_2$	$O \rightarrow O\text{-}Dep \rightarrow H \leftarrow F\text{-}Dep \leftarrow F$ s.t. $F \in \{F\}$, $O/F\text{-}Dep \in \{MR\}$, $POS(O) \in \{JJ\}$	$o = O$	same as R1$_2$ with *iPod* is the known word and *best* as the extract word.
R3$_1$	$F_{i(j)} \rightarrow F_{i(j)}\text{-}Dep \rightarrow F_{j(i)}$ s.t. $F_{j(i)} \in \{F\}$, $F_{i(j)}\text{-}Dep \in \{CONJ\}$, $POS(F_{i(j)}) \in \{NN\}$	$f = F_{i(j)}$	*Does the player play dvd with audio and "video"?* *video*→*conj*→*audio*
R3$_2$	$F_i \rightarrow F_i\text{-}Dep \rightarrow H \leftarrow F_j\text{-}Dep \leftarrow F_j$ s.t. $F_i \in \{F\}$, $F_i\text{-}Dep == F_j\text{-}Dep$, $POS(F_j) \in \{NN\}$	$f = F_j$	*Canon "G3" has a great len.* *len*→*obj*→*has*←*subj*←*G3*
R4$_1$	$O_{i(j)} \rightarrow O_{i(j)}\text{-}Dep \rightarrow O_{j(i)}$ s.t. $O_{j(i)} \in \{O\}$, $O_{i(j)}\text{-}Dep \in \{CONJ\}$, $POS(O_{i(j)}) \in \{JJ\}$	$o = O_{i(j)}$	*The camera is amazing and "easy" to use.* *easy*→*conj*→*amazing*
R4$_2$	$O_i \rightarrow O_i\text{-}Dep \rightarrow H \leftarrow O_j\text{-}Dep \leftarrow O_j$ s.t. $O_i \in \{O\}$, $O_i\text{-}Dep == O_j\text{-}Dep$, $POS(O_j) \in \{JJ\}$	$o = O_j$	*If you want to buy a sexy, "cool", accessory-available mp3 player, you can choose iPod.* *sexy*→*mod*→*player*←*mod*←*cool*

Identify opinion orientation

- For each feature, we identify the sentiment or opinion orientation expressed by a reviewer.
- Almost all approaches make use of opinion words and phrases. But notice again (a simplistic way):
  - Some opinion words have context independent orientations, e.g., “great”.
  - Some other opinion words have context dependent orientations, e.g., “small”
  - Many ways to use opinion words.
- Machine learning methods for sentiment classification at the sentence and clause levels are also applicable.

Aggregation of opinion words

(Hu and Liu, KDD-04; Ding and Liu, 2008)

- Input: a pair (*f*, *s*), where *f* is a product feature and *s* is a sentence that contains *f*.
- Output: whether the opinion on *f* in *s* is positive, negative, or neutral.
- Two steps:
  - Step 1: split the sentence if needed based on BUT words (but, except that, etc).
  - Step 2: work on the segment s_f containing *f*. Let the set of opinion words in s_f be w_1, .., w_n. Sum up their orientations (1, -1, 0), and assign the orientation to (*f*, *s*) accordingly.
- In (Ding et al, WSDM-08), step 2 is changed to $$\sum_{i=1}^{n} \frac{w_i.o}{d(w_i, f)}$$ with better results. $w_i.o$ is the opinion orientation of w_i. $d(w_i, f)$ is the distance from *f* to w_i.

Basic Opinion Rules (Liu, Ch. in NLP handbook)

Opinions are governed by some rules, e.g.,

1. Neg → Negative
2. Pos → Positive
3. Negation Neg → Positive
4. Negation Pos → Negative
5. Desired value range → Positive
6. Below or above the desired value range → Negative

Basic Opinion Rules (Liu, Ch. in NLP handbook)

7. Decreased Neg → Positive
8. Decreased Pos → Negative
9. Increased Neg → Negative
10. Increased Pos → Positive
11. Consume resource → Negative
12. Produce resource → Positive
13. Consume waste → Positive
14. Produce waste → Negative

Divide and Conquer

- Most current techniques seem to assume one-technique-fit-all solution. Unlikely??
  - "The picture quality of this camera is great."
  - "Sony cameras take better pictures than Nikon".
  - "If you are looking for a camera with great picture quality, buy Sony."
  - "If Sony makes good cameras, I will buy one."
- Narayanan, et al (2009) took a divide and conquer approach to study conditional sentences

Roadmap

- Opinion mining – problem definition
- Document level sentiment classification
- Sentence level sentiment classification
- Opinion lexicon generation
- Feature-based opinion mining
- **Opinion mining of comparative sentences**
- Opinion spam detection
- Summary

Two Main Types of Opinions

- Direct Opinions: direct sentiment expressions on some target objects, e.g., products, events, topics, persons.
  - E.g., “the picture quality of this camera is great.”
- Comparative Opinions: Comparisons expressing similarities or differences of more than one object. Usually stating an ordering or preference.
  - E.g., “car x is cheaper than car y.”

Comparative Opinions (Jindal and Liu, 2006)

- *Gradable*
  - *Non-Equal Gradable*: Relations of the type *greater* or *less than*
    - *Ex: "optics of camera A is better than that of camera B"*
  - *Equative*: Relations of the type *equal to*
    - Ex: "*camera A and camera B both come in 7MP*"
  - *Superlative*: Relations of the type *greater* or *less than all others*
    - Ex: "*camera A is the cheapest camera available in market*"

Types of comparatives: non-gradable

- Non-Gradable: Sentences that compare features of two or more objects, but do not grade them. Sentences which imply:
  - Object A is similar to or different from Object B with regard to some features.
  - Object A has feature F_1, Object B has feature F_2 (F_1 and F_2 are usually substitutable).
  - Object A has feature F, but object B does not have.

Mining Comparative Opinions

- Objective: Given an opinionated document *d*,. Extract comparative opinions:

  (O_1, O_2, *F*, *po*, *h*, *t*),

  where O_1 and O_2 are the object sets being compared based on their shared features *F*, *po* is the preferred object set of the opinion holder *h*, and *t* is the time when the comparative opinion is expressed.

- Note: not positive or negative opinions.

Roadmap

- Opinion mining – problem definition
- Document level sentiment classification
- Sentence level sentiment classification
- Opinion lexicon generation
- Feature-based opinion mining
- Comparative opinion mining
- **Opinion spam detection**
- Summary

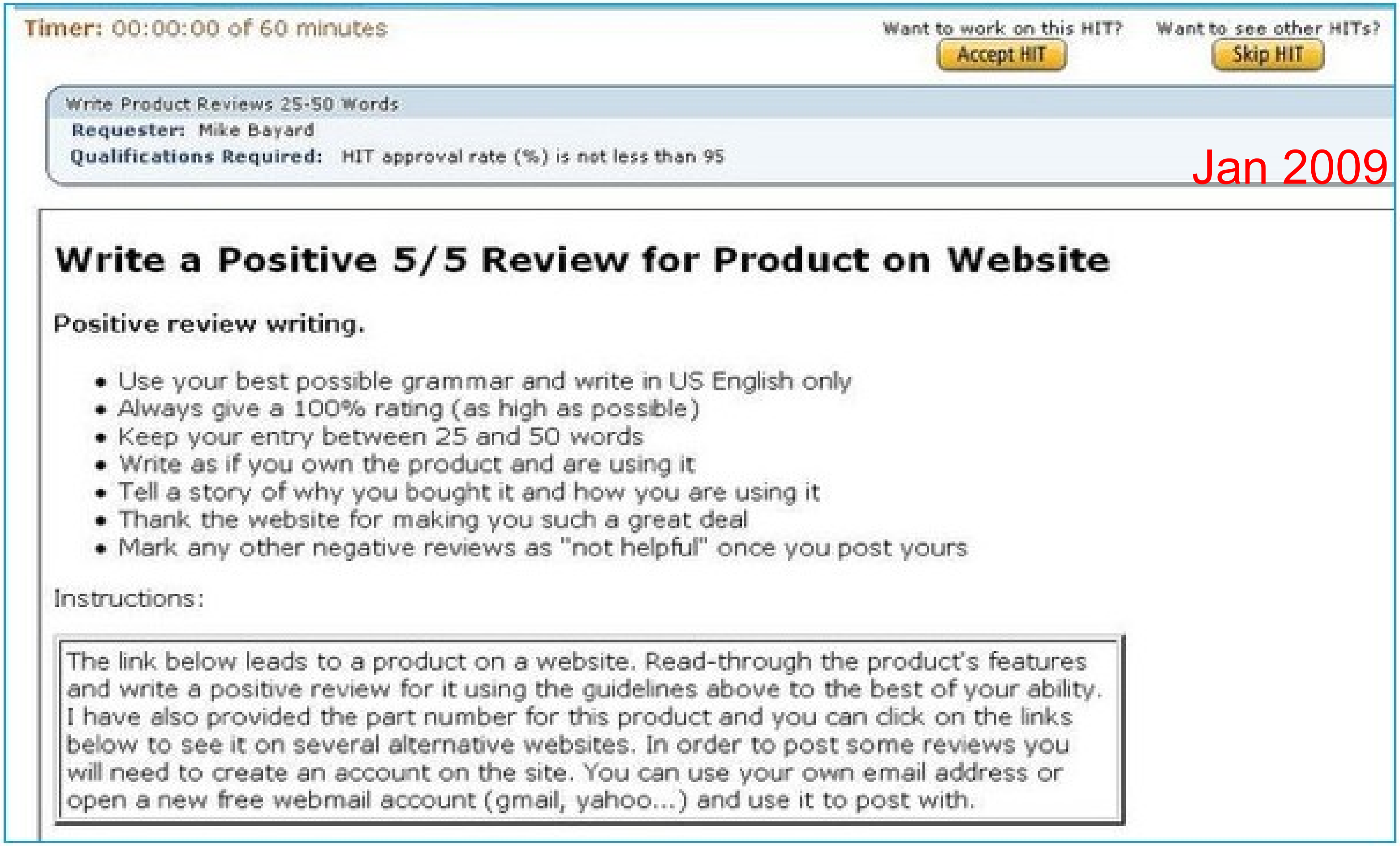
Opinion Spam Detection (Jindal and Liu, 2007)

- Fake/untruthful reviews:
  - Write undeserving positive reviews for some target objects in order to promote them.
  - Write unfair or malicious negative reviews for some target objects to damage their reputations.
- Increasing number of customers wary of fake reviews (biased reviews, paid reviews)

An Example of Practice of Review Spam

Belkin International, Inc

- Top networking and peripherals manufacturer | Sales ~ $500 million in 2008
- Posted an ad for writing fake reviews on amazon.com (65 cents per review)



Timer: 00:00:00 of 60 minutes

Write Product Reviews 25-50 Words

Requester: Mike Bayard

Qualifications Required: HIT approval rate (%) is not less than 95

Jan 2009

Write a Positive 5/5 Review for Product on Website

Positive review writing.

- Use your best possible grammar and write in US English only
- Always give a 100% rating (as high as possible)
- Keep your entry between 25 and 50 words
- Write as if you own the product and are using it
- Tell a story of why you bought it and how you are using it
- Thank the website for making you such a great deal
- Mark any other negative reviews as "not helpful" once you post yours

Instructions:

The link below leads to a product on a website. Read-through the product's features and write a positive review for it using the guidelines above to the best of your ability. I have also provided the part number for this product and you can click on the links below to see it on several alternative websites. In order to post some reviews you will need to create an account on the site. You can use your own email address or open a new free webmail account (gmail, yahoo...) and use it to post with.

Experiments with Amazon Reviews

- June 2006
  - 5.8mil reviews, 1.2mil products and 2.1mil reviewers.

- A review has 8 parts
  - *<Product ID> <Reviewer ID> <Rating> <Date> <Review Title> <Review Body> <Number of Helpful feedbacks> <Number of Feedbacks> <Number of Helpful Feedbacks>*

- Industry manufactured products "*mProducts*"

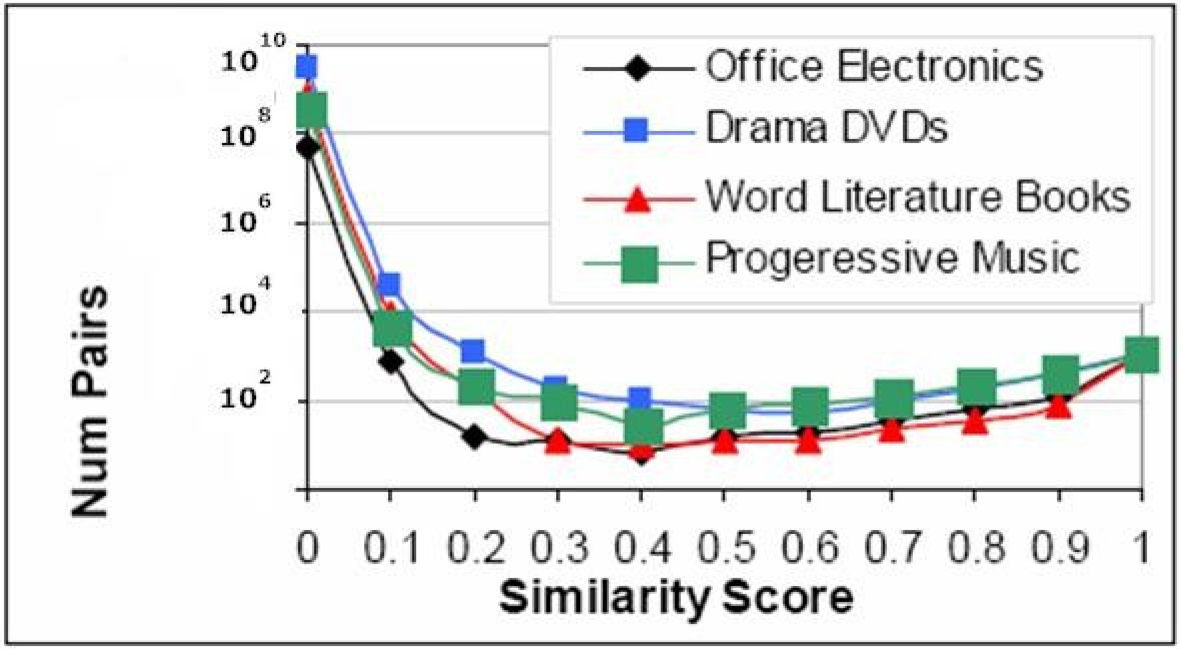
  e.g. electronics, computers, accessories, etc
  - 228K reviews, 36K products and 165K reviewers.

Deal with fake/untruthful reviews

- We have a problem: because
  - It is extremely hard to recognize or label fake/untruthful reviews manually.
  - Without training data, we cannot do supervised learning.
- Possible solution:
  - Can we make use certain duplicate reviews as fake reviews (which are almost certainly untruthful)?

Duplicate Reviews

Two reviews which have similar contents are called duplicates



Four types of duplicates

1. Same userid, same product
2. Different userid, same product
3. Same userid, different products
4. Different userid, different products

- The last three types are very likely to be fake!

Supervised model building

- Logistic regression
  - Training: duplicates as spam reviews (positive) and the rest as non-spam reviews (negative)
- Use the follow data attributes
  - Review centric features (content)
    - Features about reviews
  - Reviewer centric features
    - Features about the reviewers
  - Product centric features
    - Features about products reviewed.

Predictive Power of Duplicates

- Representative of all kinds of spam
- Only 3% duplicates accidental
- Duplicates as positive examples, rest of the reviews as negative examples

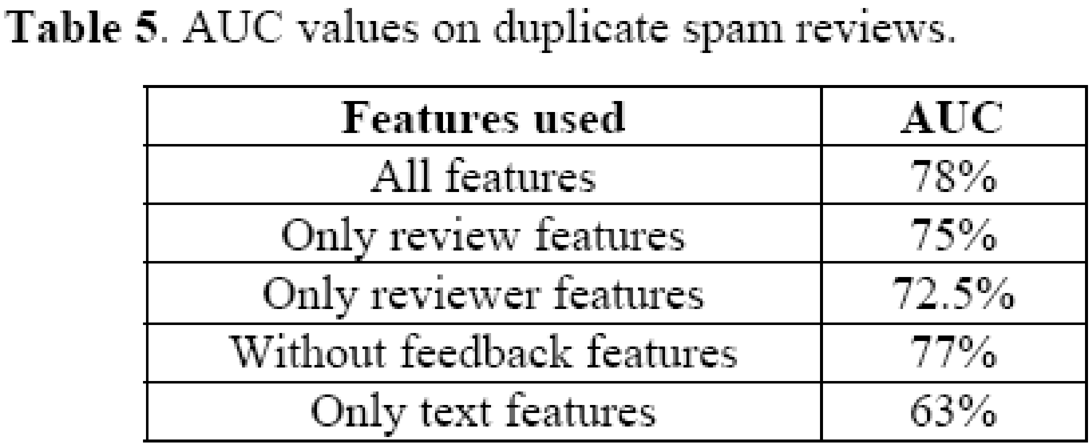


Table 5. AUC values on duplicate spam reviews.

Features used	**AUC**
All features	78%
Only review features	75%
Only reviewer features	72.5%
Without feedback features	77%
Only text features	63%

  - reasonable predictive power
  - Maybe we can use duplicates as type 1 spam reviews(?)

Spam Reviews

- Hype spam – promote one's own products
- Defaming spam – defame one's competitors' products

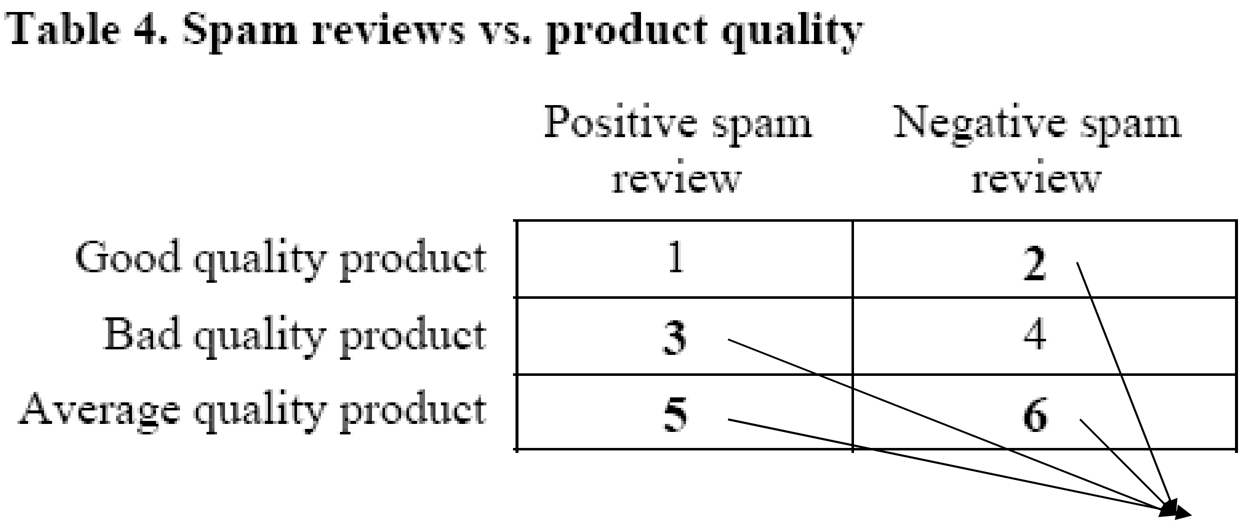


Table 4. Spam reviews vs. product quality

	Positive spam review	Negative spam review
Good quality product	1	**2**
Bad quality product	**3**	4
Average quality product	**5**	**6**

- Harmful Regions

Harmful Spam are Outlier Reviews?

- Outliers reviews:
    - Reviews which deviate from average product rating
- Harmful spam reviews:
- Outliers - necessary, but not sufficient, condition for harmful spam reviews.

Some Tentative Results

- Negative outlier reviews tend to be heavily spammed.
- Those reviews that are the only reviews of some products are likely to be spammed
- Top-ranked reviewers are more likely to be spammers
- Spam reviews can get good helpful feedbacks and non-spam reviews can get bad feedbacks

Roadmap

- Opinion mining – problem definition
- Document level sentiment classification
- Sentence level sentiment classification
- Opinion lexicon generation
- Feature-based opinion mining
- Opinion mining of comparative sentences
- Opinion spam detection
- **Summary**

Summary

- We briefly defined and introduced
  - Direct opinions: document, sentence and feature level
  - Comparative opinions: different types of comparisons
  - Opinion spam detection: fake reviews.
- There are already many applications.
- Technical challenges are still huge.
  - Accuracy of all tasks is still a major issue
- But I am optimistic. Accurate solutions will be out in the next few years. Maybe it already there.
  - A lot of unknown methods from industry.