Opinion Mining

Feiyu XU & Xiwen CHENG

Xiwen.cheng@dfki.de
DFKI, Saarbruecken, Germany
Dec 3, 2007

Outline

Introduction – Opinion Mining

- What is an opinion?
  - **[Quirk et al., 1985]**

    Private state: a state that is not open to objective observation or verification
  - **Wikipedia**

    a person's ideas and thoughts towards something. It is an assessment, judgment or evaluation of something. An opinion is not a *fact*, because opinions are either not falsifiable, or the opinion has not been proven or verified. If it later *becomes* proven or verified, it is no longer an opinion, but a fact. Accordingly, all information on the web, from a surfer's perspective, is better described as opinion rather than fact.

- Opinion Mining (OM)
  - A recent discipline at the crossroads of information retrieval, text mining and computational linguistics which tries to detect the opinions expressed in the natural language texts.
  - Opinion Extraction is a specified method of information extraction, delivering inputs for opinion mining
  - Sentiment analysis and sentiment classification are sub-areas of opinion extraction and opinion mining

Introduction – Linguistic Phenomena [Polanyi & Zaenen, 2006, TA van Dijk, 1995]

- Simple Valence
  - E.g. lexicalized, “good” (positive) & “bad” (negative)
- Prior & Contextual Polarity
  - Philip Clap, President of the National Environment ~~Trust~~, sums up well the general thrust of the reaction of environmental movements: there is no reason at all to believe that the ~~polluters~~ are suddenly going to become reasonable.

Introduction – Linguistic Phenomena (cont.)

- Contextual Valence Shifters
  - Sentence Based Contextual Valence Shifters
    - E.g. *John is successful at tennis*.
    - E.g. *John is never successful at tennis*.
    - E.g. *Rather efficient*, *deeply suspicious*
  - Modals
    - E.g. *Mary is a terrible person. She is mean to her dogs.*
    - E.g. *If Mary were a terrible person, she would be mean to her dogs.*
  - Presuppositional Items
    - E.g. *It is sufficient*.
    - *E.g. It is barely sufficient*.
    - E.g. *We want a fancy look and feel*.
    - E.g. *It would be nice if we could have the curved shape*.

Introduction – Linguistic Phenomena (cont.)

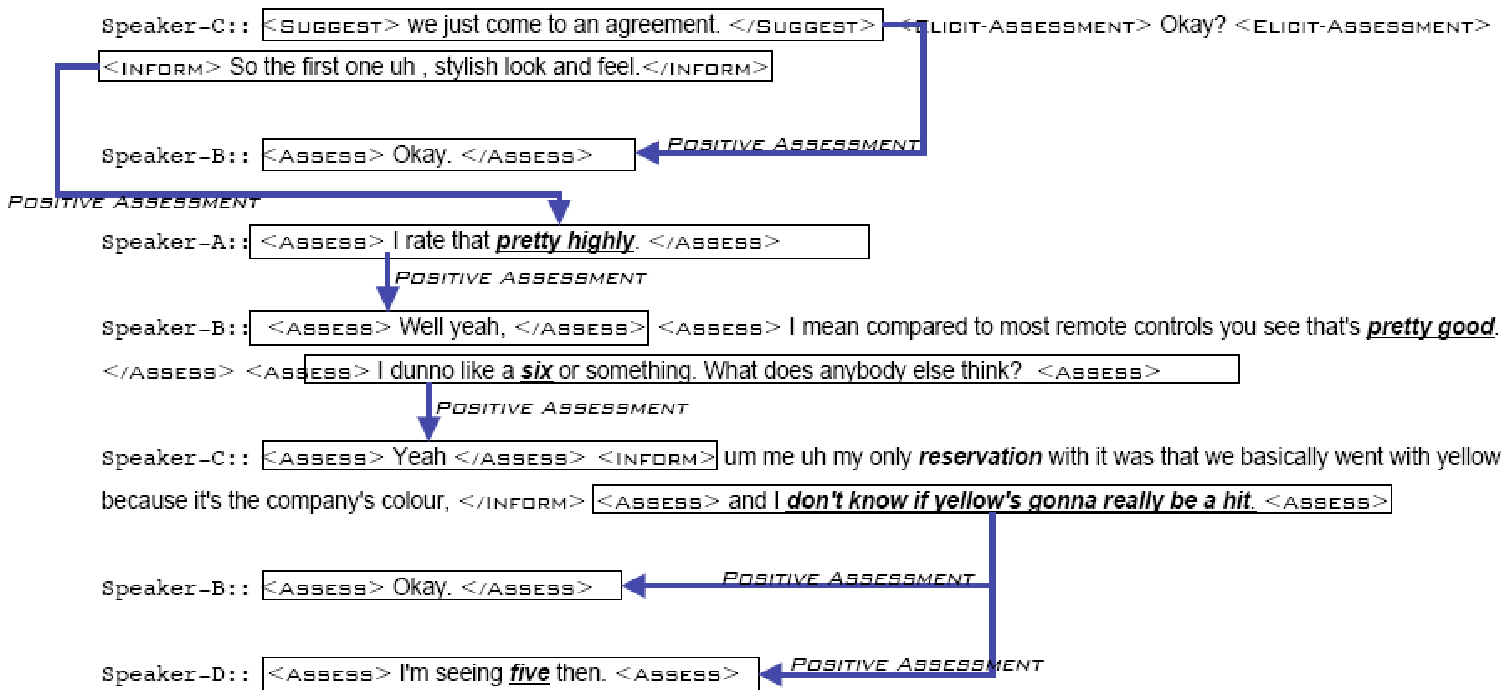
- Contextual Valence Shifters
  - Tense
    - E.g. *This is my favorable car*.
    - E.g. *This was my favorable car*.
  - Collocation
    - E.g. *It is expensive*. (about prize)
    - E.g. *It looks expensive*. (about appearance)
  - Irony
    - E.g. *The very brilliant organizer failed to solve the problem*.
    - E.g. *Terrorists deserve no mercy!*

Introduction – Linguistic Phenomena (cont.)

- Discourse Based Contextual Valence Shifters
  - Connectors
    - E.g. Although Boris is brilliant at math, he is a horrible teacher.
  - Discourse Structure
    - Lists and elaborations
    - E.g. The 7 Series is a large, well-furnished luxury sedan. The iDrive control system, which uses a single knob to control the audio, navigation, and phone systems, is meant to streamline the cabin, but causes frustration. A midcycle freshening brought revised styling, a 4.8-liter, 360-hp V8, and a new name: the 750i. The six-speed automatic shifts smoothly.
  - Multi-entity Evaluation
    - E.g. *Coffee is expensive, but Tea is cheap*.
  - Comparative
    - E.g. *In market capital, Intel is way ahead of AMD*.

Introduction – Linguistic Phenomena (cont.)

- Discourse Based Contextual Valence Shifters
  - Genre and Attitude Assessment
    - E.g. AMI meeting snippet where the participants rate their TV remote control design on a number of metrics such as learnability, look and feel, etc, using a scale from one (worse) to seven (best).



Introduction – Linguistic Phenomena (cont.)

- Discourse Based Contextual Valence Shifters
  - Reported Speech
    - E.g. *Mary was a slob*. Vs. *John said that Mary was a slob*.
  - Subtopics
    - E.g. The economic situation is more than satisfactory. The leading indicators show a rosy picture. When one looks at the human rights picture, one is struck by the increase in arbitrary arrests, by needless persecution of helpless citizens and increase of police brutality.
  - Genre Constraints
    - E.g. This film should be brilliant. The characters are appealing. Stallone plays a happy, wonderful man. His sweet wife is beautiful and adores him. He has a fascinating gift for living life fully. It sounds like a great story, however, the film is a failure.

Introduction – Applications [Liu, 2007]

- Market Intelligence: product, event and service benchmarking
  - Consumer opinion summarization
    - E.g. Which groups among our customers are unsatisfied? Why?
  - Public opinion identification and direction
    - E.g. What are the opinions of the Americans about the European style cars?
  - Recommendation
    - E.g. New Beetles is the favorite car of the young ladies.
  - Consultants
  - Virtual sale experts
  - Marketing predication

- Opinion retrieval / search
  - Opinion-oriented search engine
  - Opinion-based question answering
    - E.g. What is the general opinion on the proposed tax reform?
  - Sentiment-enhanced machine translation

Outline

- Introduction
  - Opinion Mining
  - Linguistic Perspective
  - Application
- Opinion Mining
  - Abstraction
  - Linguistic Resource of OM
  - Document, Sentence, Clause Level Sentiment Analysis
  - Feature-based Opinion Mining and Summarization
  - Comparative Sentence and Relation Extraction
- Conclusion
  - Resource
  - Challenges

Opinion Mining – Abstraction [Liu, Web Data Mining book 2007]

- Basic components of an opinion
  - Opinion holder: The person or organization that holds a specific opinion on a particular object
  - Object: on which an opinion is expressed
  - Opinion: a view, attitude, or appraisal on an object from an opinion holder
  - E.g. *John said that Mary was a slob*.
- Object/Entity: An object is an entity which can be a product, person, event, organization, topic, or even opinion. It can be represented as an ontology including
  - a hierarchy of concepts and their sub-concepts, where
  - each concept can be associated with a set of attributes or properties,
  - In practice, we often use "feature" to represent both concepts and attributes.
  - E.g. Gas mileage of VW Golf is great !
    - Domain: Car; Instance: VW Golf; Attribute: mileage
- Opinion: An opinion can be expressed on any ontology node or attribute of the node. The opinion could be 3-ary or scalable.
  - *Positive, Negative, Neutral*
  - *-10 -- +10*

Opinion Mining – Model of a review [Liu, Web Data Mining book 2007]

- An object O is represented with a finite set of features, $F=\{f_1, f_2, \ldots, f_n\}$
  - Each feature f_i in F can be expressed with a finite set of words or phrases W_i, E.g. *V-6, V-8 → Engine*
  - Another word, we have a set of corresponding synonym sets $W=\{W_1, W_2, \ldots, W_n\}$ for the features

- Model of a review: An opinion holder j comments on a subset of the features $S_j \subseteq F$ of object O
  - For each feature $f_k \in S_j$ that j comments on, he/ she
    - Chooses a word or phrase from W_k to describe the feature, and
    - Expresses a positive, negative or neutral opinion on f_k
  - E.g. Expensive design & Expensive prize

OM – Research topics

- Development of linguistic resources for OM
  - Automatically build lexicons of subjective terms
- At the document/sentence/clause level
  - Assumption: each document, sentence or clause focuses on a single object and contains opinion (positive, negative and neutral) from a single opinion holder
  - Subjective / objective classification
  - Sentiment classification: positive, negative and neutral
  - Strength Detection of opinions from clauses
- At the feature level
  - Identify and extract commented features
  - Group feature synonyms
  - Determine the sentiments towards these features
- Comparative opinion mining
  - Identify comparative sentences
  - Extract comparative relations from these sentences

OM – Linguistic Resource of OM [Esuli, 2006]

- Linguistic resource of OM are **opinion words or phrases** which are used as instruments for sentiment analysis. It also called polar words, opinion bearing words, subjective element, etc.

- Research word on this topic deal with three main tasks:
  - Determining term **orientation**, as in deciding if a given Subjective term has a Positive or a Negative slant
  - Determining term **subjectivity**, as in deciding whether a given term has a Subjective or an Objective (i.e. neutral, or factual) nature.
  - Determining the **strength** of term attitude (either orientation or subjectivity), as in attributing to terms (real-valued) degrees of positivity or negativity.

- Example
  - Positive terms: *good, excellent, best*
  - Negative terms: *bad, wrong, worst*
  - Objective terms: *vertical, yellow, liquid*

Orientation of terms [Esuli, 2006]

The problem:

Determining if a subjective term has a Positive or a Negative orientation.

[Hatzivassiloglou and McKeown, 1997]

Hypothesis: *adjectives* in and conjunctions usually have similar orientation, though but is used with opposite orientation.

Example (conjuction of adjectives)

1. The tax proposal was *simple and well received...*
2. The tax proposal was *simplistic but well received...*
3. * The tax proposal was *simplistic and well received...*

Method: a weighted graph of similarity of orientation is defined by analyzing conjunctions of adjectives in unprocessed text, then a minimum-cut method is applied to the graph.

Orientation of terms [Esuli, 2006]

[Turney and Littman, 2003]

Hypothesis: terms with similar orientation tend to co-occur in documents.

The *Semantic Orientation* (SO) of a term is estimated by combining a pointwise mutual information (PMI) measure of the term against some *paradigmatic* terms.

Pos = {`good`, `nice`, `excellent`, `positive`, `fortunate`, `correct`, `superior`}
Neg = {`bad`, `nasty`, `poor`, `negative`, `unfortunate`, `wrong`, `inferior`}

PMI is measured using the number of results returned by the AltaVista search engine.

$$PMI(t, t_i) = \log \frac{\#(\text{“}t \texttt{ NEAR } t_i\text{”})}{\#(\text{“}t\text{”})\#(\text{“}t_i\text{”})}$$

$$SO(t) = \Sigma_{t_i \in Pos} PMI(t, t_i) - \Sigma_{t_i \in Neg} PMI(t, t_i)$$

OM – Linguistic Resource of OM

- Advantage:
  - Naive solution to achieve prior polarities
- Exception:
  - Subjective & Objective
    - E.g. *long & excellent*
  - Conjunction
    - E.g. Nice and Nasty (the first hit from Google for “Nice and *”)
  - Contextual polarity
    - E.g. It looks cheap. It is cheap.
    - E.g. It is expensive. It looks expensive.
- Term with more than one senses
  - Different senses of the same ambiguous term may have different sentiment-related properties
  - Example:
    - Estimable – ambiguous term with an objective sense (i.e. measurable), and a positive sense (i.e. deserving respect)

Orientation of terms [Esuli, 2006]

[Esuli and Sebastiani, 2005]

Hypothesis: terms with similar orientation have similar glosses.

Example (glosses for terms with similar orientation)

`good`: "*that which is pleasing or valuable or useful*"; "*agreeable or pleasing*".

`beautiful`: "*aesthetically pleasing*".

`pretty`: "*pleasing by delicacy or grace; not imposing*".

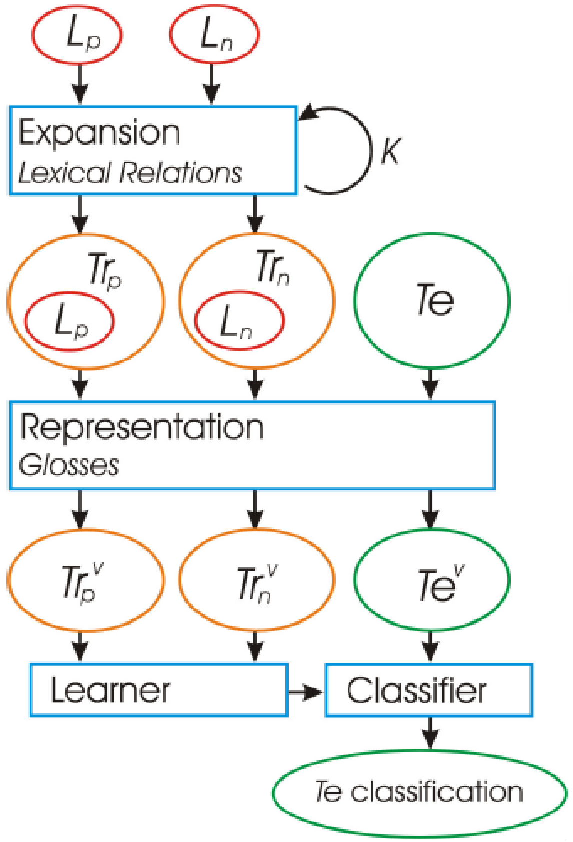
Each term is represented by its gloss.

A binary classifier is learned, in a semi-supervised process, using the glosses of the Positive and Negative terms in the training set.

Orientation of terms

A semi-supervised learning method to determine semantic orientation of terms:

- The training set is built by iteratively adding to it *synonyms* and *antonyms* of terms already belonging to it, starting from two small seed sets L_p and L_n of known *Positive* and *Negative* terms.
- A classifier is learned on the glosses of terms in training set and then applied to the glosses of terms in test set.



Orientation of terms [Esuli, 2006]

Test sets:

HM: 657 Positive / 679 Negative hand labeled adjectives, defined in [Hatzivassiloglou and McKeown, 1997].

TL: 1,614/1,982 terms extracted from the General Inquirer (GI) lexicon.

Results:

Test set	Method	Accuracy(%)
HM	[Hatzivassiloglou and McKeown, 1997]	78.08
	[Turney and Littman, 2003] AV-NEAR	87.13
	[Turney and Littman, 2003] 7M-NEAR	80.31
	[Esuli and Sebastiani, 2005]	**87.38**
TL	[Turney and Littman, 2003] AV-NEAR	82.84
	[Turney and Littman, 2003] 7M-NEAR	76.06
	[Turney and Littman, 2003] AV-AND	67.00
	[Esuli and Sebastiani, 2005]	**83.09**

Subjectivity of terms [Esuli, 2006]

[Esuli and Sebastiani, 2006a]

The method of [Esuli and Sebastiani, 2005] is adapted to classify terms as either Positive, Negative or Objective.

Hypothesis:

- (from previous work) terms with similar orientation have similar glosses.
- terms without orientation have *non-oriented* glosses.

Example

`yellow`: "*similar to the color of an egg yolk*".

`vertical`: "*at right angles to the plane of the horizon or a base line*".

Test set: the whole GI lexicon (1,614 Pos/1,982 Neg/5,009 Obj).

Results: 67.6% accuracy on classification on Subjective vs Objective, 66.0% on classification on the three categories.

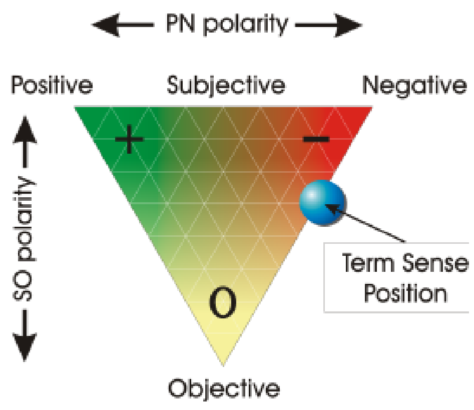
Subjectivity and orientation of term senses [Esuli, 2006]

[Esuli and Sebastiani, 2006b]
Previous experiences on terms showed that:

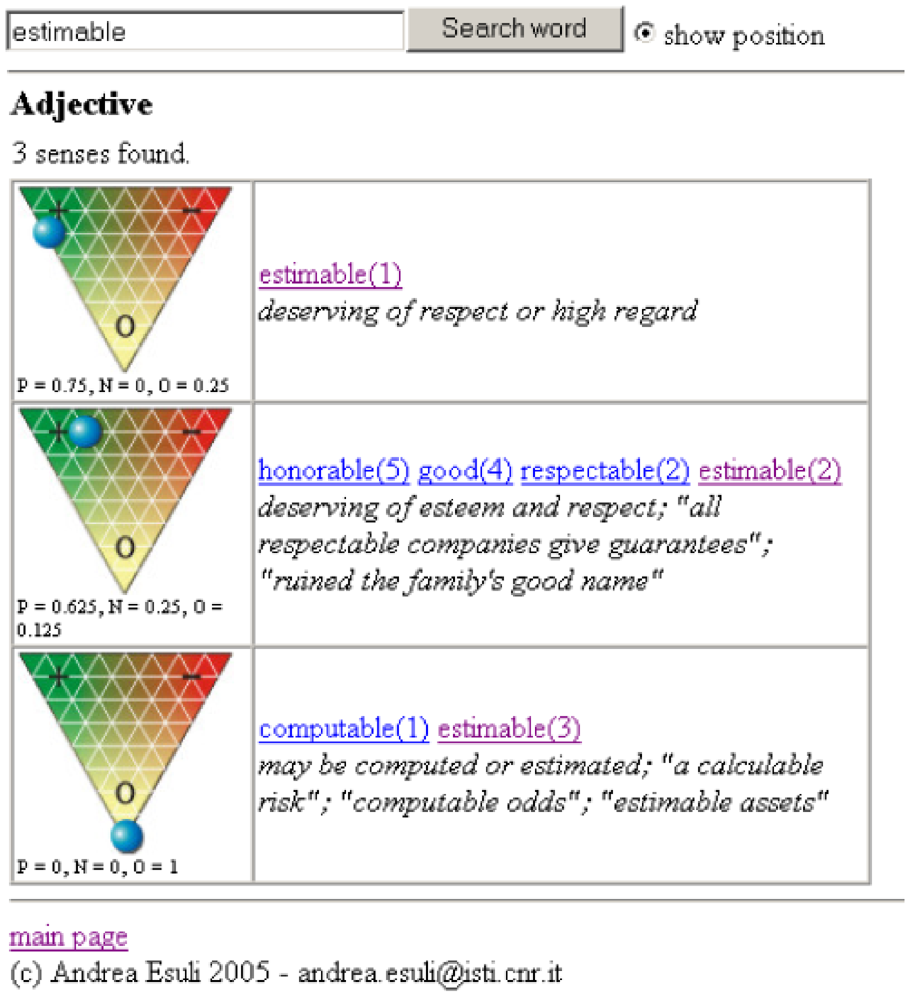
- Variation in the parameters of the classifiers do not affect accuracy but *distribution of terms among categories*.
- "Diffult" terms are those that have multiple senses with different sentiment properties (e.g. `bright`, `high`).

The method of [Esuli and Sebastiani, 2006a] has been adapted to classify each synset of WordNet, using various configuration of the classifier.

SentiWordNet is a lexical resource that assigns to each synset of WordNet three sentiment scores: positivity, negativity, objectivity.
The sum of the scores for a synset is always 1



[LREC’06] SENTIWORDNET interface [Esuli, 2006]



OM – Linguistic Resource of OM [Esuli, 2006]

- Advantage:
  - Dimensional polarity
- Could be improved:
  - Usage for contextual polarity
- Exception:
  - Multi-word expressions
    - Not entirely satisfactory – negative expression
    - *"There is __no reason at all to believe__ it is the best car."*

Subjectivity properties of multi-word expressions: The Appraisal Theory

[Martin and White, 2005] – The Appraisal theory.

Appraisal theory is a framework of linguistic resources which describe how writers and speakers express inter-subjective and ideological positions.

- *appraisal*
  - *attitude*
    - *affect*
    - *appreciation*
    - *judgement*
  - *graduation*
    - *force*
    - *focus*
  - *orientation*
    - *positive*
    - *negative*
  - *polarity*
    - *marked*
    - *unmarked*

	happy	very	"very happy"	not	"not very happy"
attitude:	affect	–	affect	–	affect
orientation:	positive	–	positive	**negate**	**negative**
force:	neutral	**increase**	**high**	**reverse**	**low**
focus:	neutral	–	neutral	–	neutral
polarity:	unmarked	–	unmarked	**marked**	**marked**

[Whitelaw et al., 2005] semi-automatically have produced a lexicon of 1,329 appraisal entities from 400 seed terms, in around twenty man-hours.

OM – Research topics

- Development of linguistic resources for OM
  - Automatically build lexicons of subjective terms
- At the document/sentence/clause level
  - Assumption: each document, sentence or clause focuses on a single object and contains opinion (positive, negative and neutral) from a single opinion holder
  - Subjective / objective classification
  - Sentiment classification: positive, negative and neutral
  - Strength Detection of opinions from clauses
  - * Less information, more challenges
- At the feature level
  - Identify and extract commented features
  - Determine the sentiments towards these features
  - Group feature synonyms
- Comparative opinion mining
  - Identify comparative sentences
  - Extract comparative relations from these sentences

OM – Document Level Sentiment Analysis

- The problem: Determining the overall sentiment properties of a text
- Unsupervised review classification
  - Turyney, 2003
- Sentiment classification using machine learning methods
  - Pang et al., 2002, Pang and Lee, 2004, Whitelaw et al., 2005
- Review classification by scoring features
  - Dave, Lawrence and Pennock, 2005

Unsupervised review classification

- Hypothesis: the orientation of the whole document is the sum of the orientation of all its parts
- Three steps
  - POS Tagging and Two consecutive word extraction (e.g. JJ NN)
  - Semantic orientation estimation (AltaVisata near operator)
    - Pointwise mutual information

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

    - Semantic orientation

      SO(phrase) = PMI(phrase, “excellent”) – PMI(phrase, “poor”)
  - Average SO Computation of all phrases
    - The review is recommended if average SO is positive, not recommended otherwise
- The average accuracy on 410 reviews is 74%, ranging from 84% for automobile reviews to 66% for movie reviews

Sentiment classification using machine learning methods

- [Pang et al., 2002]
  - Apply some standard supervised automatic text classification methods to classify orientation of movie reviews
    - Learners: Naive Bayes, MaxEnt, SVM
    - Features: unigrams, bigrams, adjective, POS, position
    - Preprocessing: negation propagation
    - Representation: binary, frequency
  - 82.9% accuracy, on a 10-fold cross validation experiments on 1,400 movie reviews (best from SVM, unigrams, binary)
- [Pang and Lee, 2004]
  - A sentence subjectivity classifier is applied, as preprocessing, to reviews, to filter out Objective sentences.
  - Accuracy on movie reviews classification raises to 86.4%
- [Whitelaw et al. 2005]
  - Appraisal features are added to the Movie Review Corpus, which obtained a 90.2% classification accuracy.

Review classification by scoring features

- Feature selection:
  - Comparing the performance from metadata and statistical substitution, linguistic substitution, language-based modification, N-grams and substring
  - Bigram is the best

- Learning algorithm
  - Comparing with NB, ME, SVM …
  - Score the features: C and C' are classes

$$score(f_i) = \frac{P(f_i \mid C) - P(f_i \mid C')}{P(f_i \mid C) + P(f_i \mid C')}$$

  - Classification a review

$$class(d_j) = \begin{cases} C & eval(d_j) > 0 \\ C' & eval(d_j) < 0 \end{cases}$$

$$eval(d_j) = \sum_i score(f_i)$$

- Accuracy 88%

OM – Document-level Sentiment Classification

- Advantage:
  - Coarse Analysis
- Exception:
  - This film should be brilliant. The characters are appealing. Stallone plays a happy, wonderful man. His sweet wife is beautiful and adores him. He has a fascinating gift for living life fully. It sounds like a great story, however, the film is a failure.

OM – Sentence Level Sentiment Analysis

- Harder than document-level sentiment classification

- [Yu and Hatzivassiloglou, 2003]: overall SO properties in sentences
  - Subjectivity & objectivity classification
    - Sentiment similarity – based on shared words and phrases, and WordNet synsets
    - Naive Bayesian classification – 1,2,3-grams, POS and opinion words
    - Multi-NB classification – based on the subsets of the above experiment
  - Sentiment orientation classification: similar as Turney, 2002
    - More seeds (1,336)
    - The average per word Log-likelihood scores

$$L(W_i, POS_j) = \log\left(\frac{\frac{Freq(W_i, POS_j, ADJ_p)+\epsilon}{Freq(W_{\text{all}}, POS_j, ADJ_p)}}{\frac{Freq(W_i, POS_j, ADJ_n)+\epsilon}{Freq(W_{\text{all}}, POS_j, ADJ_n)}}\right)$$

OM – Sentence-level Sentiment Classification

- Advantage:
  - Even though the analysis is still coarse, it is more specific than document-level analysis
- Exception:
  - *The very brilliant organizer failed to solve the problem.*

OM – Sentence Level Sentiment Analysis (cont.)

- [Rilloff and Wiebe, 2003]: subjective / objective classification
  - Taking advantages of Information Extraction techniques
  - Manually collected opinion words + AutoSlog-TS

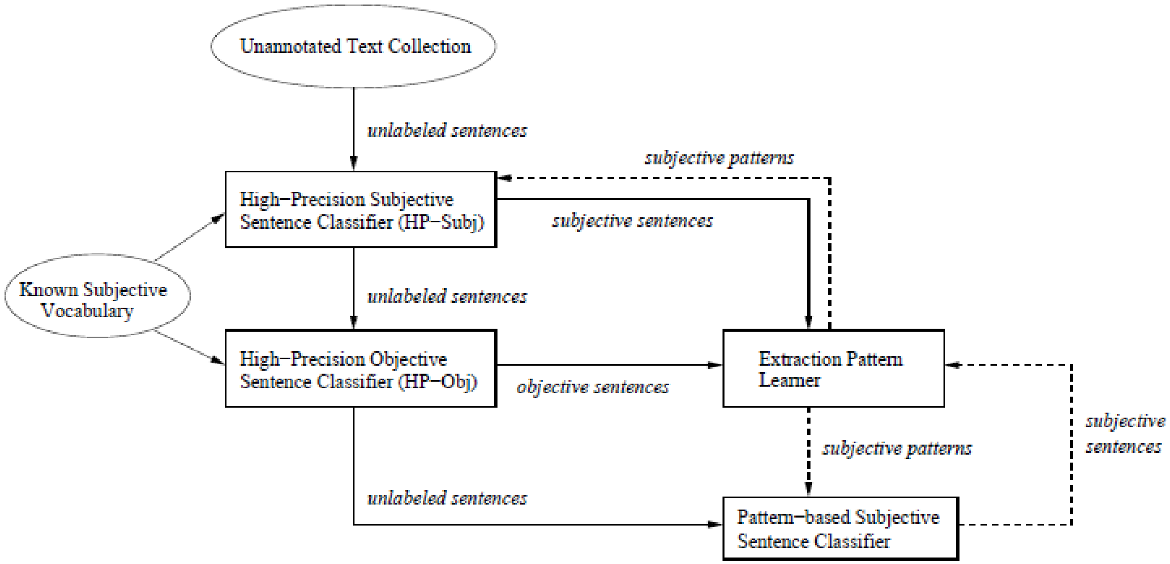


Figure 1: Bootstrapping Process

<subject> passive-vp	*<subj> was satisfied*
<subject> active-vp	*<subj> complained*
<subject> active-vp dobj	*<subj> dealt blow*
<subject> active-vp infinitive	*<subj> appears to be*
<subject> passive-vp infinitive	*<subj> was thought to be*
<subject> auxiliary dobj	*<subj> has position*

active-vp <dobj>	*endorsed <dobj>*
infinitive <dobj>	*to condemn <dobj>*
active-vp infinitive <dobj>	*get to know <dobj>*
passive-vp infinitive <dobj>	*was meant to show <dobj>*
subject auxiliary <dobj>	*fact is <dobj>*

passive-vp prep <np>	*opinion on <np>*
active-vp prep <np>	*agrees with <np>*
infinitive prep <np>	*was worried about <np>*
noun prep <np>	*to resort to <np>*

OM – Sentence-level Sentiment Classification

- Advantage:
  - *The very brilliant organizer failed to solve the problem.*
    - *Very brilliant organizer → subj*
    - *Subj failed*
- Exception:
  - *It sounds like a great story, however, the film is a failure.*

OM – Clause Level Sentiment Analysis [Wilson et al., 2006]

- The problem: automatic classification of the intensity of opinions being expressed in clauses
  - Clause: based on the non-leaf verbs in the dependency tree
  - Intensity: e.g. good, very good; bad, pretty bad
- Supervised machine learning
  - Annotation:

    President Mohammad Khatami of Iran, whose attempt at reforms have gotten American <low>support</>, <high>accused</> the United States of "<high>warmongering</>."
- Feature selection and organization
  - Type and Intensity
- Learning package and Result
  - BoosTexter, Ripper, SVMlight
  - Using SVM, mean-squared error ranging from 49% to 51%
  - Using boosting, accuracy ranging from 23% to 96%

Feature selection and organization

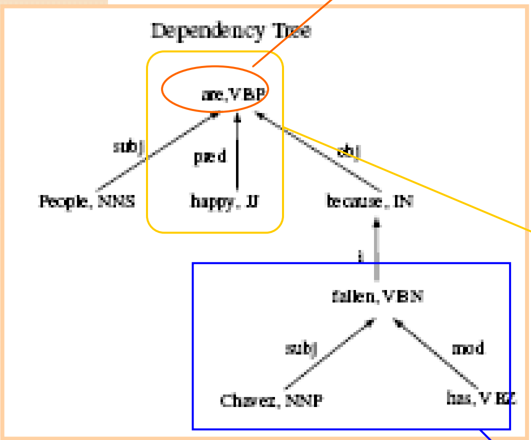
- Lexicon clues
  - Verbs of judgment (e.g. commend, reprove), desire (e.g. fancy, pine), and psych (e.g. dread, love) from Levin's (1993) English verb classes
  - Words and phrases culled from Ballment and Brennenstuhl's (1981) speech act verb classes (e.g. advocate, grumble about)
  - Verbs and adjectives listed in FrameNet (Baker, Fillmore, and Lowe, 1998) with frame element *experiencer* (e.g. Emotion_active (fuss, worry), Experiencer_obj(embarrass, thrill), Perception_body(ache, tickle)…).
  - Adjectives manually annotated for polarity from (Hatzivassiloglou and McKeown, 1997) (e.g. positive (appealing, brilliant…), negative (bizarre, dismal…)).
  - Subjectivity clues listed in (Wiebe, 1990) (e.g. absurdly, funny)

Feature selection and organization

- Lexicon clues
  - Annotated data: Distributionally similar adjectives and verbs and n-grams (Wiebe et al., 2004)
    - E.g. *worst of all, of the century, do something about, on the other hand, price you have to, etc.*
  - Unannotated data: extraction patterns and subjective nouns (Rillof and Wiebe, 2003; Riloff, Wiebe, and Wilson, 2003)
    - E.g. *<subj> was hired, <subj> dealt blow*
  - Low-frequency words (Wiebe et al., 2004)
    - E.g. *bleat and bore, womanize and booze, so enthusiastic, so cubersome, etc.*

Feature selection and organization

- Syntax clues



root

root-lex(w, t): word w with POS tag t is the root of a dependency tree (i.e., the main verb of the sentence).

root-backoff(t): a word with POS tag t is the root of a dependency tree.

leaf

leaf-lex(w, t): word w with POS tag t is a leaf in a dependency tree (i.e., it has no modifiers).

leaf-backoff(t): a word with POS tag t is a leaf in a dependency tree

node

node-lex(w, t): word w with POS tag t.

node-backoff(t): a word with POS tag t.

bilex

bilex-lex(w, t, r, w_c, t_c): word w with POS tag t is modified by word w_c with POS tag t_c, and the grammatical relationship between them is r.

bilex-backoff(t, r, t_c): a word with POS tag t is modified by a word with POS tag t_c, and the grammatical relationship between them is r.

allkids

allkids-lex$(w, t, r_1, w_1, t_1, \ldots, r_n, w_n, t_n)$: word w with POS tag t has n children. Each child word w_i has POS tag t_i and modifies w with grammatical relationship r_i, where $1 \leq i \leq n$.

allkids-backoff$(t, r_1, t_1, \ldots, r_n, t_n)$: a word with POS tag t has n children. The i^{th} child word has POS tag t_i and modifies the parent word with grammatical relationship r_i.

OM – Clause-level Sentiment Classification

- Advantage:
  - Fine-grained analysis
- Exception:
  - Feature-based Analysis
    - *The price is really cheap.*
    - *The quality is really cheap.*

OM – Research topics

- Development of linguistic resources for OM
  - Automatically build lexicons of subjective terms
- At the document/sentence/clause level
  - Assumption: each document, sentence or clause focuses on a single object and contains opinion (positive, negative and neutral) from a single opinion holder
  - Subjective / objective classification
  - Sentiment classification: positive, negative and neutral
  - Strength Detection of opinions from clauses
  - * **Less information, more challenges**
- At the feature level
  - Identify and extract commented features
  - Group feature synonyms
  - Determine the sentiments towards these features
- Comparative opinion mining
  - Identify comparative sentences
  - Extract comparative relations from these sentences

OM – Feature-based OM and Summarization [Hu and Liu, 2004]

GREAT Camera., Jun 3, 2004
Reviewer: **jprice174** from Atlanta, Ga.

I did a lot of research last year before I bought this camera... It kinda hurt to leave behind my beloved nikon 35mm SLR, but I was going to Italy, and I needed something smaller, and digital.

The pictures coming out of this camera are amazing. The 'auto' feature takes great pictures most of the time. And with digital, you're not wasting film if the picture doesn't come out. …

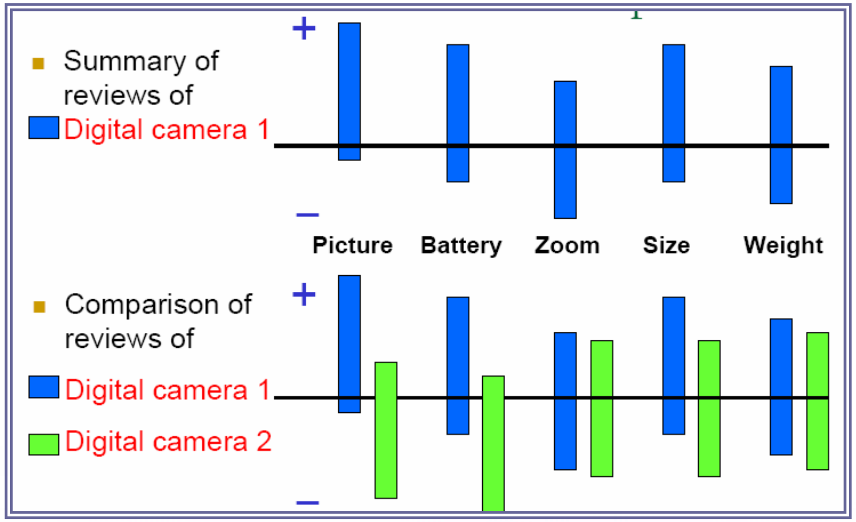
Feature1: picture
Positive: 12
- The pictures coming out of this camera are amazing.
- Overall this is a good camera with a really good picture clarity.

…

Negative: 2
- The pictures come out hazy if your hands shake even for a moment during the entire process of taking a picture.
- Focusing on a display rack about 20 feet away in a brightly lit room during day time, pictures produced by this camera were blurry and in a shade of orange.

Prior & contextual SO
- *E.g. Hotel Review:*
  - *hot water*
  - *hot room*
- *E.g. Car Review*
  - *looks expensive*
  - *Is expensive*

Feature extraction:
- Explicit & Implicit
  - E.g. great photos <photo>
  - E.g. small to keep <size>
- Frequent & Infrequent



Featured-based – Feature Extraction

- Frequent & Infrequent features
  - Frequent feature: <u>Label sequential rules</u>
    - Annotation
      - *"Included memory is stingy"*
      - <{included, VB}{$feature, NN}{is, VB}{stingy, JJ}>
    - Learned LSRs
      - <{easy, JJ}{to}{*, VB}>→ <{easy, JJ}{to}{$feature, VB}>
      - [sup=10%, conf=95%]
    - Feature extraction
      - The word in the sentence segment of a new review that matches $feature is extracted
  - Infrequent feature
    - Observation: the same *opinion word* can be used to describe different features and objects
      - E.g. The pictures (high-freq) are absolutely amazing.
      - E.g. The software (low-freq) that comes with it is amazing.

Label sequential rules

- Label sequential rule (LSR) : X → Y
  - Y is a sequence and X is a sequence produced from Y by replacing some of its items with wildcards
  - A data instance (S_i, y_i) in D is said to *cover* the LSR if X is a subsequence of S_i
  - A data instance (S_i, y_i) is said to *satisfy* a LSR if Y is a subsequence of S_i.
  - The support (sup) of the rule is the fraction of total instances in D that satisfies the rule.
  - The confidence (conf) of the rule is the proportion of instances in D that covers the rule also satisfied the rule.

- Example
  - <{1}{3}{*,*}>→ <{1}{3}{7,8}> [sup=2/5, conf=2/3]

	Data Sequence
1	⟨{1}{3}{5}{7, 8, 9}⟩
2	⟨{1}{3}{6}{7, 8}⟩
3	⟨{1, 6}{9}⟩
4	⟨{1}{3}{5, 6}⟩
5	⟨{1}{3}{4}⟩

Featured-based – Group Feature Synonyms

- Identify part-of relationship [Popescu and Etziono, 2005]
  - Each noun phrase is given a PMI score with part discriminators (e.g. of scanner, scanner has) associated with the product class, (e.g. a scanner class)
- Liu et al., 2005 use WordNet
- Carenini et al., 2005 is based on similarity metrics
  - The system merges each discovered feature to a feature node in the pre-set taxonomy
  - The similarity metrics are defined based on string similarity, synonyms and other distances measured using WordNet

- Camera
  - Lens
    - Digital Zoom
    - Optical Zoom
    - ...
  - Editing/Viewing
    - Viewfi nder
    - ...
  - Flash
  - ...
- Image
  - Image Type
    - TIFF
    - JPEG
    - ...
  - Resolution
    - Effective Pixels
    - Aspect Ratio
  - ...

Feature Extraction and Group

- Advantage:
  - Explicit features
- Unsolved
  - Implicit features
  - Ontology-based Feature Identification
- Exception:
  - More relation from sentences
    - *Gas Mileage of VW Golf is great.*
      - *Entity: VW Golf*
      - *Attribute: Gas Mileage*
  - Concept Grouping: *<automobile engine version>*
    - *V12 8000CC is pretty powerful.*
    - *V6 4000CC is not a real good engine.*
  - Coverage of WordNet to identify part-of relation
    - *"2000 Honda Accord Coupe" AS car entity*

OM – Research topics

- Development of linguistic resources for OM
  - Automatically build lexicons of subjective terms
- At the document/sentence/clause level
  - Assumption: each document, sentence or clause focuses on a single object and contains opinion (positive, negative and neutral) from a single opinion holder
  - Subjective / objective classification
  - Sentiment classification: positive, negative and neutral
  - Strength Detection of opinions from clauses
  - * **Less information, more challenges**
- At the feature level
  - Identify and extract commented features
  - Group feature synonyms
  - Determine the sentiments towards these features
- Comparative opinion mining
  - Identify comparative sentences
  - Extract comparative relations from these sentences

Featured-based Sentiment Orientation [Popescu and Etzioni, 2005]

- Contextual Semantic Orientation
  - <word, SO>, <word, feature, SO>, <word, feature, sentence, SO>
    - E.g. **SEN:*"I am not happy with this sluggish driver."***
    - ***<sluggish, ?>, <sluggish, driver, ?>, <sluggish, driver, SEN, ?>***

- Relaxation labeling: sentiment assignment to words satisfying local constraints.
  - *Constraints*:
    - conjunctions, disjunctions, syntactic dependency rule, morphological relationships, WordNet-supplied synonymy and antonymy, etc.
  - *Neighborhood*: a set of words connected the word through constraints.

–Initialization: prior polarity assignment (PMI-based approach [Turney, 2003])

–Support function: The influence of an object's neighborhood on its Label

•E.g. *q("sluggish", POS)=P("sluggish/POS" | {<"happy/POS" >})*
+ P("sluggish/POS" | {<"satisfy/POS", >})…

–Update equation: Reestimate the probability of updating labels on top of constraints for each iteration

•E.g. "***hot(?) room and broken(-) fan***" → hot(-)

Featured-based Sentiment Orientation [Ding and Liu 2007]

- Prior polarity assignment
  - Hypothesis: Less influence from far away opinion words
    - Sentence segmentation using BUT words/phrases (e.g. "*but*", "*except that*", etc.)
    - In each segment, low weights assignment to opinion words that are far away from the feature.
    - E.g. "*The camera has a **long** **battery life**, which is **great**.*"
- Context-dependent Opinions: Linguistic rules or conversions
  - Intra-sentence conjunction rule
    - E.g. "*the battery life is very long(?)*"
    - E.g. "*This camera takes great(+) pictures **and** has a long battery life.*"
      - » Great (positive) → long (positive) towards "*battery life*"
  - Pseudo intra-sentence conjunction rule
    - E.g. "*The camera has a long battery life, **which** is great(+).*"
      - » Great (positive) → long (positive) towards "*battery life*"
  - Inter-sentence conjunction rule
    - Connectors between sentences, e.g. "*but*", "*however*", etc.
- Evaluation: It outperformed the pervious method around 4% of F-score.

Featured-based Sentiment Orientation

- ?
  - *"The price is* ***cheap but*** *the quality is quite* ***high****."*
  - *"The quality is* ***cheap but*** *prices are* ***expensive.****"*
  - *"But compared to conventional turbo-diesel family hatchbacks it looks* ***expensive and slightly under-equipped.*** *"*
  - *Towards a travel: "It looks* ***expensive and boring****."*

OM – Research topics

- Development of linguistic resources for OM
  - Automatically build lexicons of subjective terms
- At the document/sentence/clause level
  - Assumption: each document, sentence or clause focuses on a single object and contains opinion (positive, negative and neutral) from a single opinion holder
  - Subjective / objective classification
  - Sentiment classification: positive, negative and neutral
  - Strength Detection of opinions from clauses
  - * **Less information, more challenges**
- At the feature level
  - Identify and extract commented features
  - Group feature synonyms
  - Determine the sentiments towards these features
- Comparative opinion mining
  - Identify comparative sentences
  - Extract comparative relations from these sentences

OM – Comparative Sentence and Relation Extraction

[Jinal and Liu, SIGIR-2006]

- Linguistic Perspective
  - Comparative sentences use morphemes like
    - *More/most, -er/-est, less/least, than and as*
  - Limitations
    - Limited coverage
      - *E.g. I prefer Intel to AMD.*
    - Non-comparatives with comparative words
      - *E.g. In the context of speed, faster means better.*
- Gradable
  - Non-Equal Gradable: greater or less
    - *E.g. Optics of camera A is better than that of camera B.*
  - Equative
    - *E.g. Camera A and camera B both come in 7MP.*
  - Superlative
    - *E.g. Camera A is the cheapest camera available in market.*
- Non-gradable
  - *E.g. Object A has feature F, but object B does not have.*

OM – Comparative Sentence and Relation Extraction

- Definition: A gradable comparative relation captures the essence of a gradable comparative sentence and is represented with the following:

  (relation word, features, entity S1, entity S2, type)

  – Relation word: The keyword used to expressed a comparative relation in a sentence. *E.g. better, ahead, most, better than*
  – Features: a set of features being compared
  – Entity S1 and Entity S2: sets of entities being compared
  – Type: non-equal gradable, equative or superlative
- Example
  – *Car X has better controls than car Y.*
    - (better, controls, car X, car Y, non-equal-gradable)
  – *Car X and car Y have equal mileage.*
    - (equal, mileage, car X, car Y, equative)
  – *Car X is cheaper than both car Y and car X.*
    - (cheaper, null, car X, car Y car Z, non-equal-gradable)
  – *Company X produces a variety of cars, but still best cars come from company Y.*
    - (best, cars, company Y, null, superlative)

Identify comparative sentences

- Extract sentences which contain at least a keyword
  - recall=98%, precision=32%
  - 83 keywords
    - Words with POS tags: JJR, JJS, RBR, RBS
    - Exceptions:
      - More, less, most and least
      - Indicative words: Best, exceed, ahead, etc
      - Phrases: in the lead, on par with, etc
- Use a NB classifier : comparative & non-comparative
  - Attribute: class sequential rules (CSRs)
    - A rule with a sequential pattern on the left and a class label on the right of the rule
    - 13 manual rules
      - Whereas/IN, but/CC, however/RB, while/IN, though/IN, etc
  - E.g. This camera has significantly more noise at ISO 100 than the Nikon 4500.
    - *<{$entityS1,NN}{has/VBZ}{*}{more/JJB} >→ comparative*

Extract comparative relations [Jindal and Liu, AAAI-2006]

- Classify comparative sentences into: non-equal gradable, equative, and superlative
  - SVM + keywords
  - If the sentence has a particular keyword in the attribute set, the corresponding value is 1, and 0 otherwise
- Extraction of relation items
  - Extraction of features, entities and relation keywords
    - *(**relation word**, **features**, **entity S1**, **entity S2**, **type**)*
  - Assumption:
    - There is only one relation in a sequence
    - Features are nouns (e.g. exception: it is small.)
  - Three steps:
    - Sequence data generation
    - Label sequential rule (LSR) generation
    - Build a sequential cover/extractor from LSRs

Sequence data/ LSRs generation

- Label Set = {$entityS1, $entityS2, $feature}
- Distance words (n=4)
  - {l1,l2,l3,l4,r1,r2,r3,r4}
    - "li" means the distance of i to the left of the pivot
    - "ri" means the distance of I to the right of the pivot
- Special words #start and #end are used to mark the start and the end of a sentence
- LSRs generation: e.g. *<{*,NN}{VBZ}>→<{$entityS1,NN}{VBZ}>*

The comparative sentence "*Canon/NNP has/VBZ better/JJR optics/NNS*" has *$entityS1* "*Canon*" and *$feature* "*optics*".

Sequences are:

- ⟨{#start}{l1}{**$entityS1**, NNP}{r1}{has, VBZ }{r2 }{better, JJR}{r3}{$Feature, NNS}{r4}{#end}⟩
- ⟨{#start}{l4}{**$entityS1**, NNP}{l3}{has, VBZ}{l2}{better, JJR}{l1}{**$Feature**, NNS}{r1}{#end}⟩

Sequential database construction

- Step 1: Select the LSR with the highest confidence (conf). Replace the matched elements in the sentences that satisfy the rule with the labels in the rule.
- Step 2: Recalculate the confidence of each remaining rule based on the modified data from the step 1.
- Step 3: Repeat step 1 and 2 until no rule left with confidence higher than the *mincof* value.

- Example
  - <{1}{3}{*,*}>→ <{1}{3}{7,8}> [sup=2/5, conf=2/3]

	Data Sequence
1	⟨{1}{3}{5}{7, 8, 9}⟩
2	⟨{1}{3}{6}{7, 8}⟩
3	⟨{1, 6}{9}⟩
4	⟨{1}{3}{5, 6}⟩
5	⟨{1}{3}{4}⟩

Performance

- Identifying gradable comparative sentences
  - Precision = 82%, recall = 81%

- Classification into three gradable types
  - SVM achieves the best result: 96%

- Extraction of comparative relations
  - LSR : F-score=72%

- Comments
  - Not all comparison are evaluations.
    - *E.g. Car X is 2 feet longer than Car Y.*
    - *E.g. Cellphone X has Bluetooth, but cellphone Y does not have.*

OM – Research topics

- Development of linguistic resources for OM
  - Automatically build lexicons of subjective terms
- At the document/sentence/clause level
  - Assumption: each document, sentence or clause focuses on a single object and contains opinion (positive, negative and neutral) from a single opinion holder
  - Subjective / objective classification
  - Sentiment classification: positive, negative and neutral
  - Strength Detection of opinions from clauses
  - * **Less information, more challenges**
- At the feature level
  - Identify and extract commented features
  - Group feature synonyms
  - Determine the sentiments towards these features
- Comparative opinion mining
  - Identify comparative sentences
  - Extract comparative relations from these sentences

⇨ OMINE – ontology-based opinion mining system

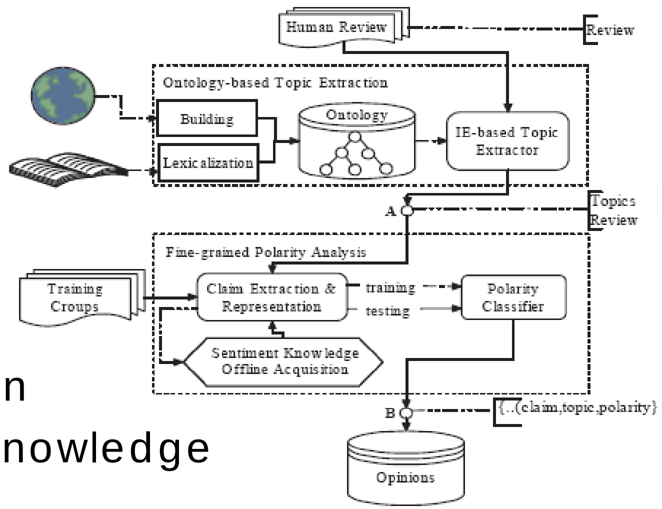
OMINE – Opinion Mining System

- Ontology-based Topic Extraction
  - E.g. “Gas mileage of VW Golf is great.”
    - Entity: VW Golf
    - Attribute: Mileage
  - E.g. <car engine>
    - *“V12 8000CC is pretty powerful.”*
    - *“V6 4000CC is not a real good engine.”*

- Fine-grained Polarity Analysis
  - IE-based approach and contextual-dependent polarity
  - Pattern is much more flexible than AutoSlog-TS
  - E.g. *“I would like more gas efficiency.”*

OMINE – Opinion Mining System

- Ontology-based Topic Extraction
  - Offline Ontology Building
  - Ontology Lexicalization
  - IE-based Topic Extraction



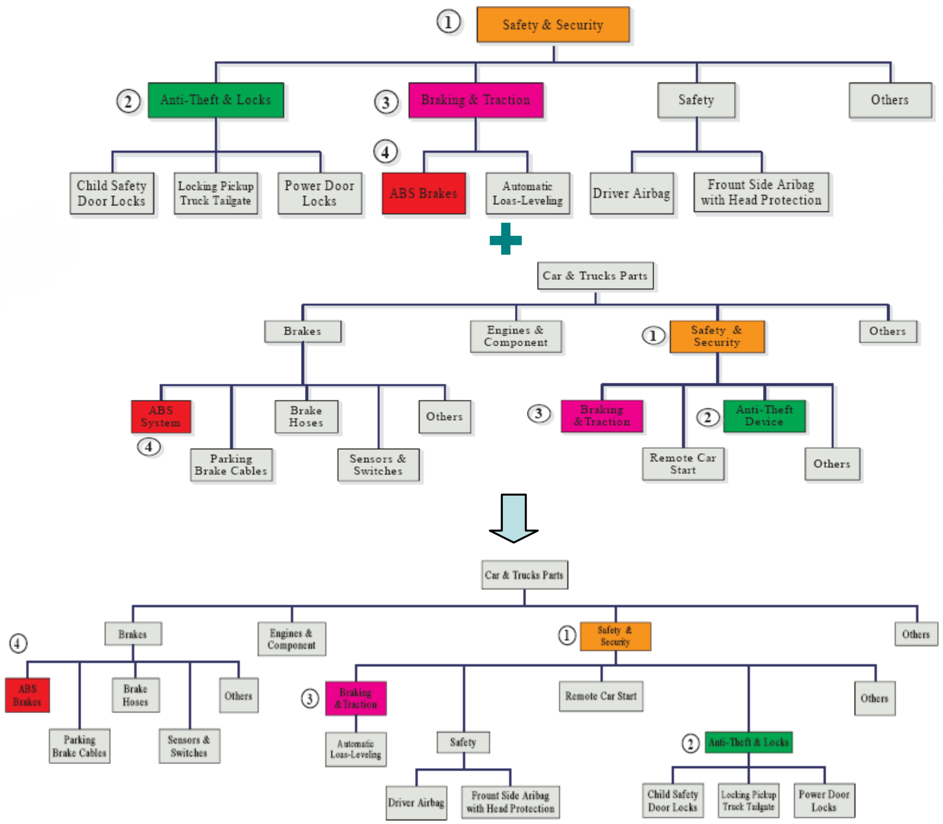
- Fine-grained Polarity Analysis
  - Claim Extraction & Representation
  - Offline Acquisition of Sentiment Knowledge
  - Polarity Analysis

Topic Extraction - Offline ontology building

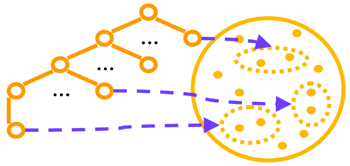
- Given a set of similar existing taxonomies, the goal is to merge arbitrary two of them iteratively until a uniform ontology is generated.
  - Concept similarity

$$CL(x,y) = \frac{|S(x) \cap S(y)|}{|S(x) \cup S(y)|}$$

  - General and Specific taxonomy Distinguishing
  - Concept pairs selection between the general and specific taxonomy
    - Maximum concept similarity
    - Maximum ST depth



Topic Extraction - Ontology Lexicalization



- In order to adapt it in real applications, the goal is to link the concepts in the ontology with everyday-used words, for instance, jargon, abbreviation, and acronym (e.g. *“transmission <-> trans”, “mileage: gas mileage, fuel mileage”).*
- Similar as ‘head-matching’-heuristic (Cimiano, et al., 2004), OMINE link the terms in an external glossary to the associated concepts.
  - Term Representation

term	
WORD	*string*
EXPLANATION	*string*
EQUATION	*list<string>*
BASICTERM	*list<string>*
FUNCTION	*list<string>*
COMPONENT	*string*

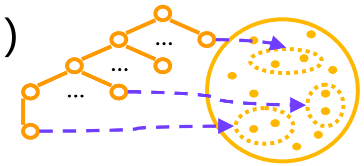
term	
WORD	*clutch pedal*
EXPLANATION	...
EQUATION	*null*
BASICTERM	*clutch, pedal*
FUNCTION	*clutch*
COMPONENT	*pedal*

- BASICTERM : atomic concepts (e.g. *engine*, *pedal, brake*)

- FUNCTION: the non-head BASICTERMs, indicates *part-of* relation (e.g. “*brake pedal*”)

- CONPONENT: the head BASICTERM, indicates *is-a* relation (e.g. “*brake pedal*”)

Topic Extraction - Ontology Lexicalization (Cont.)



- Concept Mapping
  - The concept: a single word:
    - The model queries it in *BASICTERM* to obtain both *is-a* and *part-of* relations;
  - The concept: a single word w_0 + an indicator
    - If the indicator is "*system*", the model executes query of w_0 in *COMPONENT*, while if it is "*part*", the model searches w_0 in *FUNCTION*
  - The concept: a compound word
    - *COMPONENT* + *FUNCTION*, in which the model searches the last word in *COM PONENT* and other words in *FUNCTION* (e.g. *w* = *break pad*, H = *brake lining pad*)
    - *COMPONENT* + *EXPLANATION*, in which the model searches other words in *EXPLANATION* (e.g. *w = fuel injector*, H = *injector*, *cold start injector*, *saturated switch injector*)
  - The terms in the *EQUATION* field of H are also saved as lexicons of related concepts (e.g. *w* = *check engine light*, H = *CEL*)

Topic Extraction - IE-based Topic Extraction

- It uses SProUT to recognize the set of product features and their related concept pairs.

id	pattern	example
1	*make=[car_make], model=[car_model], property=[car_property]*	*BMW, Jetta, Coupe*
2	*component=(det)?(car_component)+*	*dash light, the 1.8 turbo engine*
3	*car=(det)?(@seek(property))?[car_autoentity]*	*a hatchback car, this vehicle*
4	*car=(det|@seek(en-year))(car_make|car_model|car_property)*	*2007 Mazda CX-7*
5	*car=(@seek(en-year))?(car_make)?[car_model]*	*2000 325i*
6	*car=(@seek(car))(property)+*	*2006 Honda Accord Coupe*
7	*car=(@seek(car))(@seek(component))(property)+*	*2002 Jetta 1.8T, 2005 VW Passat TDI 4dr Wagon*

Topic Extraction - Experiment

- Data
  - Taxonomy Resource: eBay http://www.ebay.com and AutoMSN http://autos.msn.com
  - Automobile glossary: http://www.autoglossary.com, around 10,000 terms
  - Data for topic extraction: 1000 sentences from UserReview of AutoMSN
  - Golden standard: 2038 terms identified manually

- CarOnto
  - 363 concepts (e.g. Air Intake & Fuel Delivery)
  - 1233 instances (e.g. 5- speed automatic overdrive)
  - 145 values (e.g. wagon for Style, 250@5800 RPM for Horsepower)
  - 803 makes and models (e.g. BMW, Z4)
  - Ontology lexicalization is applied to 363 concepts and retrieves 9033 lexicons.
  - 11214 domain-specific lexicon instances as total

- Topic Extraction
  - TermExtractor (Sclano and Velardi, 2007)
  - OPINE (Popescu and Etzioni, 2005)

OntoTpcEx	Recall	Precision
Before Enrichment	20.97%	88.12%
After Enrichment	**89.35%**	**94.44%**
TermExtractor	15.72%	97.46%
OPINE	79.44%	93.12%

Polarity Analysis- Offline Acquisition of Semantic Knowledge

- It aims to acquire lexical semantic orientations and negation words automatically
- Semantic Orientation
  - Observation: sentiment words occur frequently in the claims which share the same polarities with their semantic orientations.

$$TFIPF(w, cat) = Card_C(w, cat) \times \log\left(\frac{n \times |\, cats \,|}{cats(w) + 1} + 1\right)$$

- Negation Word
  - Assumption: for each lexicon potential claim, if we can find another lexicon potential claim, they have opposite polarities and one of them has just one more word than the other. This word is a potential negation word

$$PoF/Po'F(w) = \frac{f_w}{f_w + \lambda f'_w + 1} - \varepsilon$$

Polarity Analysis- Polarity Classification

- Feature Representation

Sentence	*The **Jetta** is the least reliable car.*
LOP	(“be”:VBE:i)((TOPIC:N:subj)((DET:Det:det))(“car”:N:pred)((DET:Det:det)(“reliable”:ADJ:mod)((“least”:ADJ:mod))))
SenOP	(“be”:VBE:i)((TOPIC:N:subj)((DET:Det:det))(“car”:N:pred)((DET:Det:det)(PRO:ADJ:mod)(“least”:ADJ:mod))))
NegSenOP	(“be”:VBE:i)((TOPIC:N:subj)((DET:Det:det))(“car”:N:pred)((DET:Det:det)(PRO:ADJ:mod)(NEG:ADJ:mod))))

Polarity Analysis- Experiment

- Data
  - Resource: UserReview From AutoMSN
  - The polarities of these reviews have already been annotated by reviewers in two classes: pro and con.
  - Around 20 thousand sentences, and 50% of them are positive and the other 50% are negative.
  - 19600 sentences are used to train the classifier, and 200 positive and 147 negative sentences are applied as a test corpus
- Acquisition of Sentiment Knowledge

Type	**Precision**	**Num**
Sentiment word	95.0%	623
Negation word	73.8%	22

POSITIVE	NEGATIVE
awesome, cute, speedy excellent, well, standard great, strong, comfortable sporty, super, adorable	unimpressive, awful, terrible useless, tremendous, costly expensive, troublesome, tight cumbersome, ugly, squeaky

POS	**Negation words**
aux	*doesn't, didn't, wouldn't, shouldn't couldn't, don't, can't, won't*
det	*no, little, least*
mod	*never, barely, not, less*

Polarity Analysis- Experiment

- Polarity Analysis
  - Accuracy for Positive: 92%; Negative: 94% (BoW: 85%)
  - Sample pattern

pattern	example
<Det very PRO TOPIC>	*a very cute looking*
<I would like more TOPIC>	*I would like more powerful engine*
<TOPIC be NEG CON>	*the car is not reliable*
<TOPIC be CON>	*the suspension is noisy*
<TOPIC work PRO>	*heater works fine and the defroster works great, seat heaters work well*
<TOPIC NEG work PRO>	*the power locks never worked fine, windshield sprayers don't work right*
<TOPIC look expensive>	*The car looks expensive*
<TOPIC look CON>	*The car looks expensive, The car looks cheap*

- Comments
  - Pattern: More flexible and general than AutoSlog-TS
  - Disadvantages: 10 times training data than usual

Opinion Mining

- What is an opinion?
  - **[Quirk et al., 1985]**

    Private state: a state that is not open to objective observation or verification
  - **Wikipedia**

    a person's ideas and thoughts towards something. **It is an assessment, judgment or evaluation of something.** An opinion is not a *fact*, because opinions are either not falsifiable, or the opinion has not been proven or verified. If it later *becomes* proven or verified, it is no longer an opinion, but a fact. Accordingly, all information on the web, from a surfer's perspective, is better described as opinion rather than fact.

- Opinion Mining (OM)
  - A recent discipline at the crossroads of information retrieval, text mining and computational linguistics which tries to detect the opinions expressed in the natural language texts.
  - Opinion Extraction is a specified method of information extraction, delivering inputs for opinion mining
  - Sentiment analysis and sentiment classification are sub-areas of opinion extraction and opinion mining

Linguistic Phenomena [Polanyi & Zaenen, 2006, TA van Dijk, 1995]

- Simple Valence
  - ✓ E.g. lexicalized, “good” (positive) & “bad” (negative)
  - – *Resource, Syntactic constraints, Machine Learning, Co-occurrence*
- Prior & Contextual Polarity
  - ✓ Philip Clap, President of the National Environment ~~**Trust**~~, sums up **well** the general thrust of the reaction of environmental movements: there is no **reason** at all to believe that the ~~**polluters**~~ are suddenly going to become **reasonable**.
  - – *Using linguistic constraints and context as features, NER, …*

Linguistic Phenomena (cont.)

- Contextual Valence Shifters
  - Sentence Based Contextual Valence Shifters
    - E.g. *John is successful at tennis*.
    - E.g. *John is never successful at tennis*.
    - E.g. *Rather efficient*, *deeply suspicious*
    - *Using N-grams as features*
  - Modals
    - E.g. *Mary is a terrible person. She is mean to her dogs.*
    - E.g. *If Mary were a terrible person, she would be mean to her dogs.*
  - Presuppositional Items
    - E.g. *It is sufficient*.
    - *E.g. It is barely sufficient*.
    - E.g. *We want a fancy look and feel*.
    - E.g. *It would be nice if we could have the curved shape*.
    - *N-Gram or IE approach*

Linguistic Phenomena (cont.)

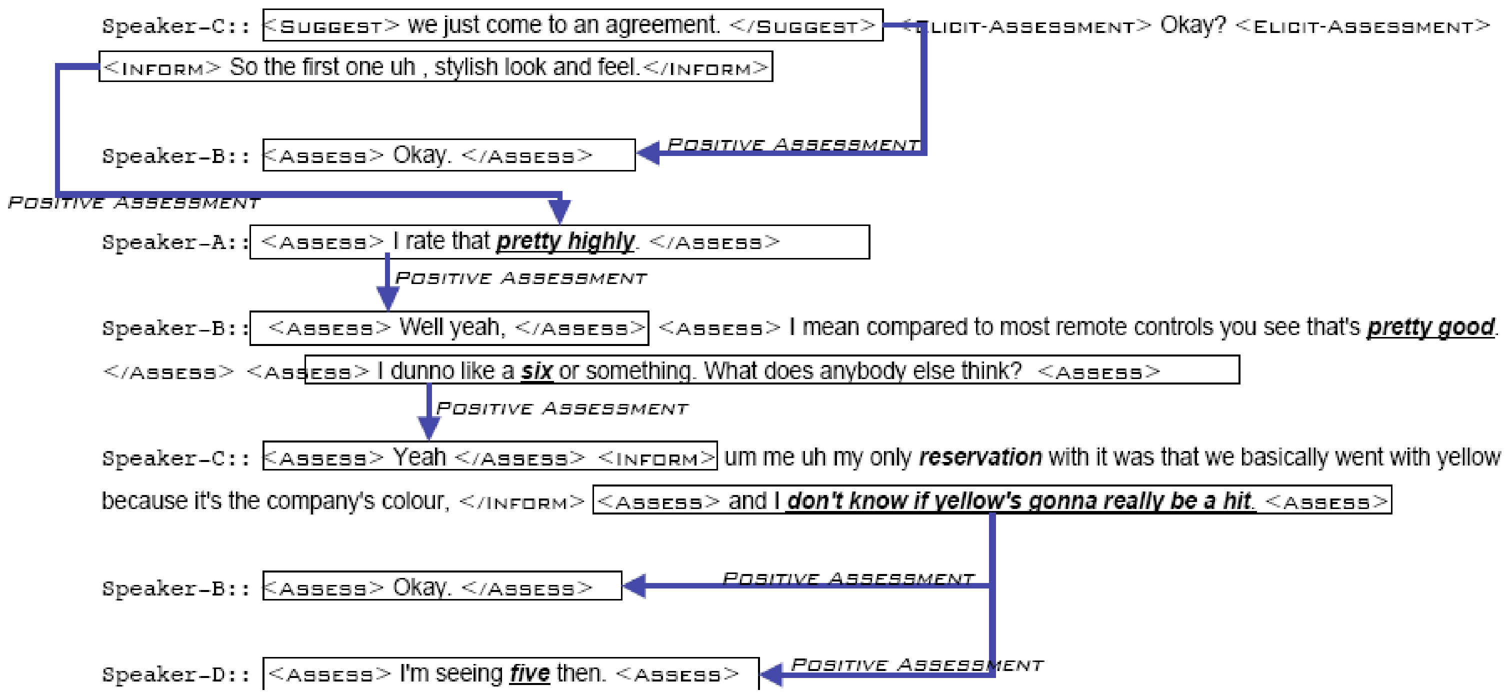
- Contextual Valence Shifters
  - ✗ Tense
    - E.g. *This is my favorable car*.
    - E.g. *This was my favorable car*.
  - ✗ Collocation
    - E.g. *It is expensive*. (about prize)
    - E.g. *It looks expensive*. (about appearance)
    - *Feature-based approach (no inexplicit FE)*
  - ✓ Irony
    - E.g. *The very brilliant organizer failed to solve the problem*.
    - E.g. *Terrorists deserve no mercy!*
    - *IE-based approach*

Linguistic Phenomena (cont.)

- Discourse Based Contextual Valence Shifters
  - ✓ Connectors
    - E.g. Although Boris is brilliant at math, he is a horrible teacher.
    - *Fine-grain*
  - ✗ Discourse Structure
    - Lists and elaborations
    - E.g. The 7 Series is a large, well-furnished luxury sedan. The iDrive control system, which uses a single knob to control the audio, navigation, and phone systems, is meant to streamline the cabin, but causes frustration. A midcycle freshening brought revised styling, a 4.8-liter, 360-hp V8, and a new name: the 750i. The six-speed automatic shifts smoothly.
  - ✓ Multi-entity Evaluation
    - E.g. *Coffee is expensive, but Tea is cheap*.
    - *Fine-grain*
  - ✓ Comparative
    - E.g. *In market capital, Intel is way ahead of AMD*.
    - *Machine learning , Feature-based approach*

Introduction – Linguistic Phenomena (cont.)

- Discourse Based Contextual Valence Shifters
  - Genre and Attitude Assessment
    - E.g. AMI meeting snippet where the participants rate their TV remote control design on a number of metrics such as learnability, look and feel, etc, using a scale from one (worse) to seven (best).



Linguistic Phenomena (cont.)

- Discourse Based Contextual Valence Shifters

Reported Speech

- E.g. *Mary was a slob*. Vs. *John said that Mary was a slob*.

Subtopics

- E.g. The economic situation is more than satisfactory. The leading indicators show a rosy picture. When one looks at the human rights picture, one is struck by the increase in arbitrary arrests, by needless persecution of helpless citizens and increase of police brutality.
- *Fine-grain , Feature-based approach*

Genre Constraints

- E.g. This film should be brilliant. The characters are appealing. Stallone plays a happy, wonderful man. His sweet wife is beautiful and adores him. He has a fascinating gift for living life fully. It sounds like a great story, however, the film is a failure.
- *Fine-grain, Feature-based approach*

Challenges

- Interaction between Pattern and Slot
  - <holder> **would like better** <object>
    - I would like better fuel mileage.
  - <object -1> drives like <**object-2**>
    - This car drives like a Porsche/a Nissan.
- Anaphoric resolution for summarization
  - E.g. "The turbo engine is a must-have, which provi acceleration."
- Others (context or semantic implication)
  - He is not the sharpest knife in the drawer.
  - She is a few fries short of a Happy Meal.
  - Stephanie McMahon is the next Stalin.
  - No one would say that John is smart.
  - My little brother could have told you that.
  - You are no Jack Kennedy.
  - They have not succeeded, and will never succeed, in breaking the will of this valiant people.
- More ...

"Who is the fairest one of all, and state your sources!"

Opinion Mining

- Basic components of an opinion
  - **Opinion holder**: The person or organization that holds a specific opinion on a particular object
  - **Object**: on which an opinion is expressed
  - **Opinion**: a view, attitude, or appraisal on an object from an opinion holder

- **Holder**: Who is the holder and what is his/her world-background
  - E.g. Age, Social status, Income, Nationality, …

- **Object/Entity**: An object is an entity which can be a product, person, event, organization, topic, or even opinion. It can be represented as an ontology including
  - a hierarchy of concepts and their sub-concepts, where
  - each concept can be associated with a set of attributes or properties,
    - E.g. This car is powerful, confident, sophisticated, a head turner, safe, roomy, durable …
    - Powerful: engine (component);
    - Safe, roomy, …: security, interior space, X (attribute of the car)

- **Opinion**: An opinion can be expressed on any ontology node or attribute of the node. The opinion could be 3-ary, scalable or others.
  - *E.g. I like this car.*
  - *E.g. It's a fast car, but I don't recommend it for soccer moms.*

Conclusion – Resource

- The Sentiment Bibliography
  http://liinwww.ira.uka.de/bibliography/Misc/Sentiment.html
- The Sentiment & Affect Yahoo! Group
  http://groups.yahoo.com/group/SentimentAI
- The General Inquirer
  http://www.wjh.harvard.edu/~inquirer
- SentiWordNet
  http://patty.isti.cnr.it/~esuli/software/SentiWordNet
- Movie Review corpus
  http://www.cs.cornell.edu/people/pabo/movie-review-data
- MPQA opinion corpus
  http://www.cs.pitt.edu/mpqa/databaserelease
- The Appraisal website
  http://grammatics.com/appraisal

Reference

- Slides
  - http://medialab.di.unipi.it/web/Language+Intelligence/OpinionMining06-06.pdf
  - http://www.cs.uic.edu/~liub/opinion-mining-and-search.pdf
- References

  - *Hatzivassiloglou, Vasileios and Kathy McKeown. 1997. Predicting the semantic orientation of adjectives. In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL-97), pages 174–181, Madrid, Spain.*
  - *Hu, Minqing and Bing Liu. 2004. Mining and summarizing customer reviews. In Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining 2004 (KDD 2004), pages 168–177, Seattle, Washington.*
  - *A. Popescu, "Extracting Product Features and Opinions from Reviews", Oren Etzioni, Proceedings of HLT-EMNLP, 2005*
  - *Wilson, Theresa, Janyce Wiebe, and Paul Hoffman. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In*
  - *Proceedings of the Human Language Technologies Conference/Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP-2005), pages 347–354, Vancouver, Canada.*
  - *X. Cheng, OMINE: Automatic Topic Term Detection and Sentiment Classification for Opinion Mining, Master thesis, 2007*
  - *…*


köszönöm
תודה!
děkuji
mahalo
고맙습니다
thank you
merci
谢谢
danke
Ευχαριστώ
شكرا
どうもありがとう
gracias