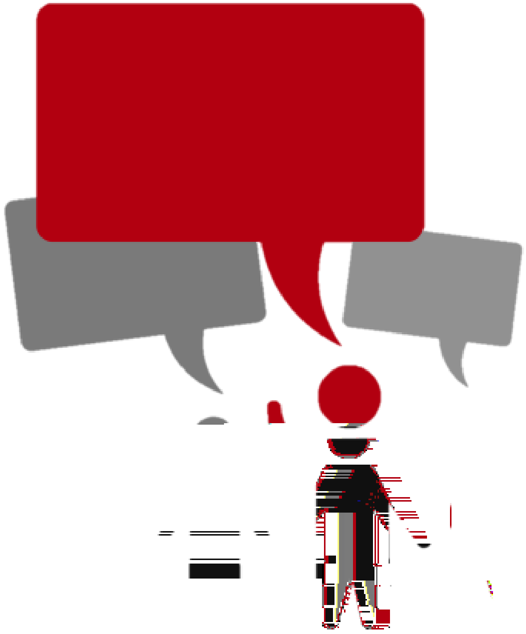
Opinion Mining



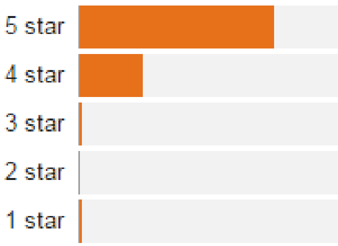
Presented by: Sinya

Outline

- Survey of Opinion Mining
- Clustering Product Features for Opinion Mining
  - Introduction
  - Related Work
  - Algorithm
  - Evaluation
  - Conclusions

Nikon D40 6.1 MP Digital SLR Camera - Black - AF-S DX 18-55mm Lens
$443 online ★★★★½ 491 product reviews

Reviews



4.7
★★★★½
491 reviews

Write a review

★★★★½ **Nikon D40 Digital Camera Review** – March 2, 2007

Patrick Singleton – Review provided by Reviewed.com Editorial review
March 2, 2007

A look at the entry-level Nikon DSLR, a 6.1-megapixel digital camera that retails for $599 with a kit lens.

★★★★☆ **Great for the price** – January 23, 2012

Stan – Review provided by Adorama
January 23, 2012

Pros: Fast / Accurate Auto-Focus; Easy To Use; Good Image Stabilization; Good Image Quality; Fast Shutter Speed...

Cons: Small LCD Screen

Great camera for the price. I use is all the time and have had no problems. Great value.

★★★★★ **"Better" than a D700** – June 8, 2009

Colin from Downunder – Review provided by Adorama
June 8, 2009

Pros: Bright LCD; Great Image Quality; Easy To Use

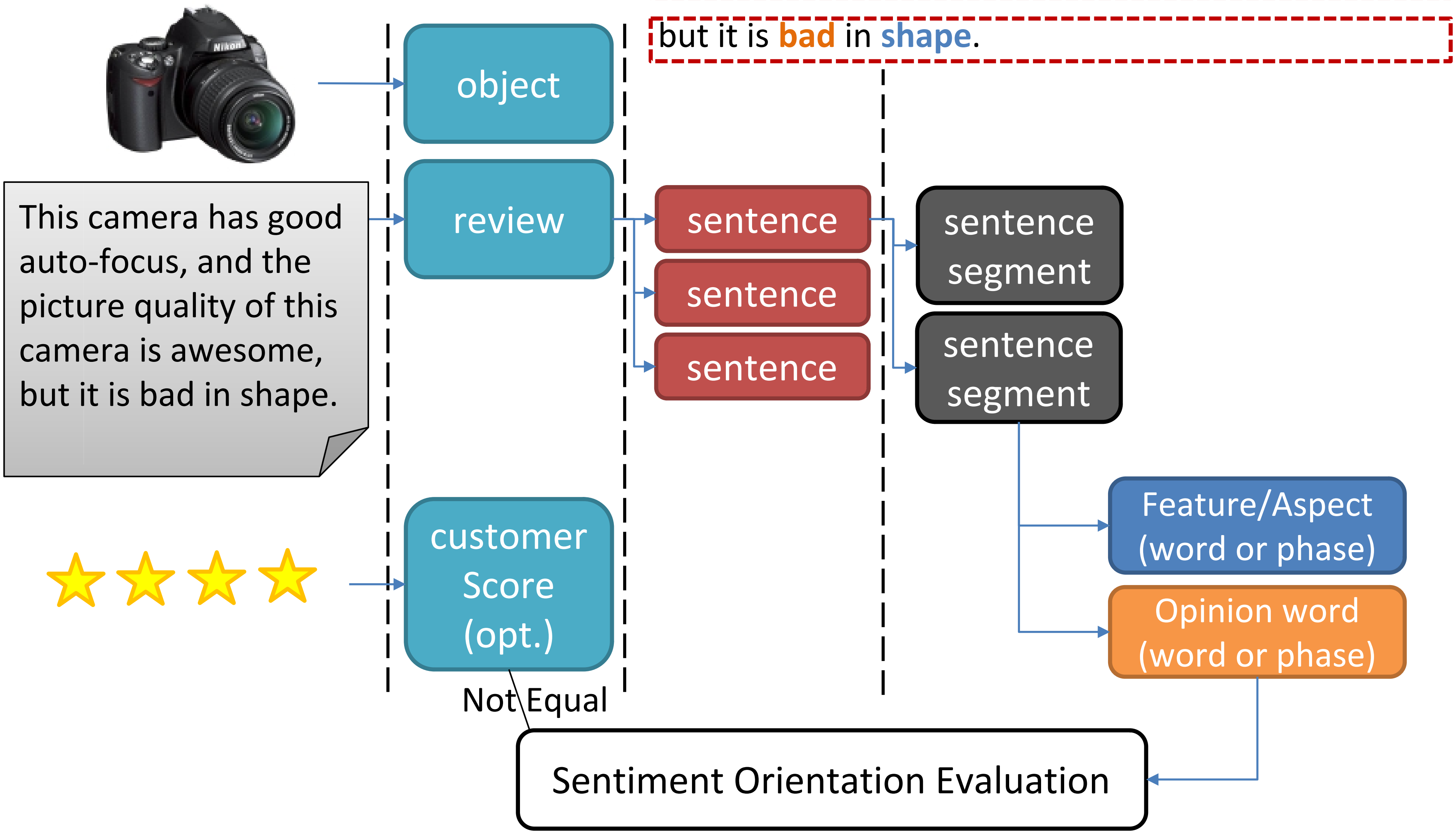
https://www.google.com/shopping/

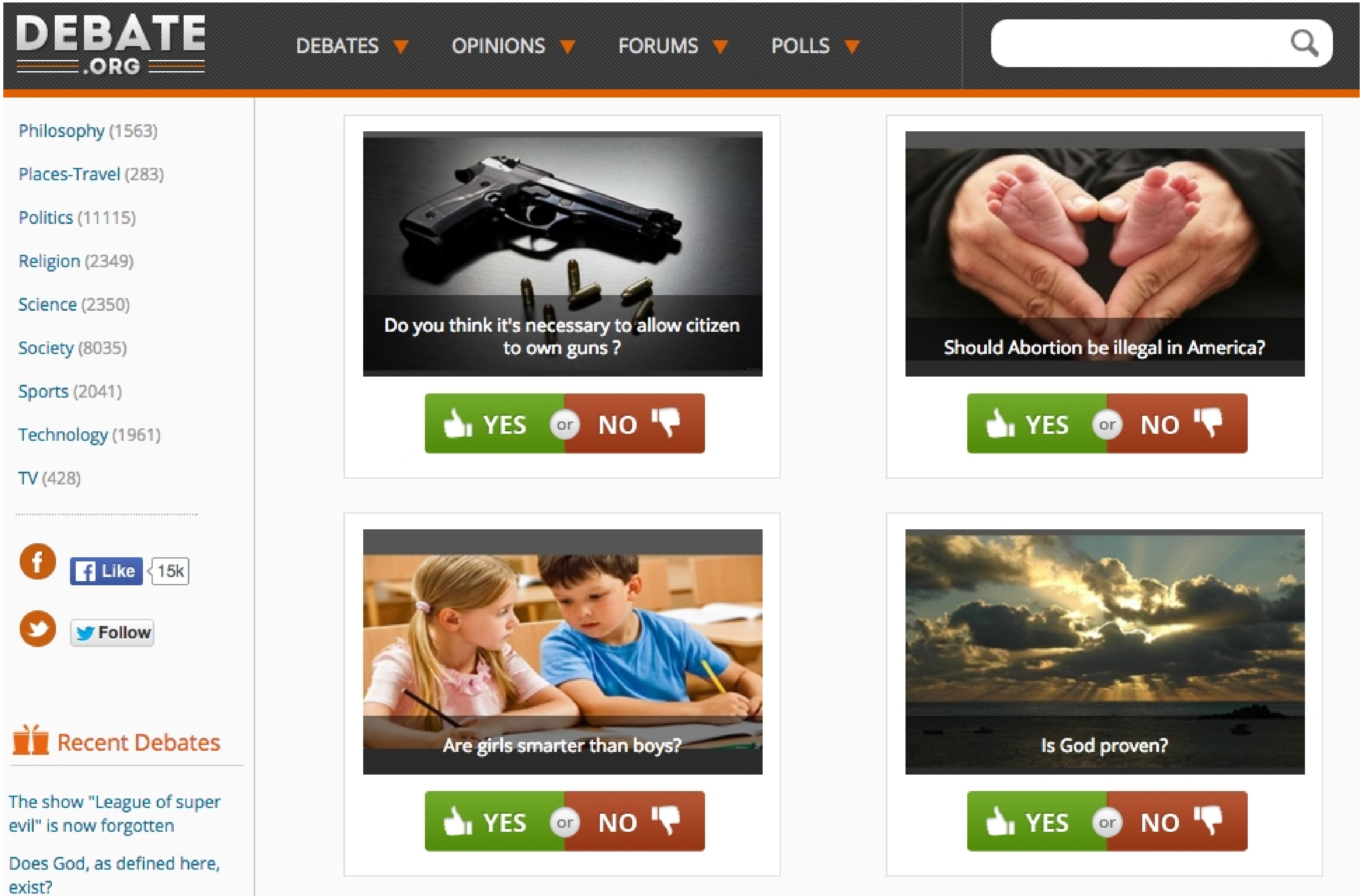
Review Model

This camera has **good** **auto-focus**,

and the **picture quality** of this camera is **awesome** …

but it is **bad** in **shape**.





http://www.debate.org/opinions/

YES or NO

Do you think it's necessary to allow citizen to own guns ?

Asked by: stephannoi

52% Say Yes

48% Say No

It's a right to own one. I don't believe anybody should be forced to own one. But my general political philosophy is that I don't believe anybody should have the right to control someone else's rights, even if it is a majority. Now is it "necessary" to own a gun? The answer to that is different for different people. Some people do go there whole lives without ever needing it. Some people needed it for self-defense. I conceal carry, and so far, I never needed a gun on the basis that I never had to use it. But that doesn't mean I never will have to use it. I can't see into the future if I'll need a gun in much the same way you can't see into the future of you needing a fire extinguisher. But you do know that it is a possibility that you might need it. So overall, the choice to own and/or carry a gun in public is a choice best left to the individual making it. Not to a democracy. It's a personal choice and a right.

Posted by: Trig314 Report Post

No, of course not, and the statistics prove it. The world's safest countries are the ones with it's gun control in check.

The murder rate in the US is the third highest out of the 36 first-world countries, behind Estonia and Latvia. It's also got a murder rate 3x higher than that of most other first-world countries.

Compare this to Australia, who 15 years ago had a severe gun problem, banned them, and has had only 1 massacre since.

You've got to be kidding me America. You're holding onto your guns so hard they're killing you.

Posted by: Kaynex Report Post

http://www.debate.org/opinions/

Roadmap

- Sentiment Analysis and Opinion Mining
  - Document sentiment classification
  - Sentence subjectivity & sentiment classification
  - Aspect-based sentiment analysis
  - Mining comparative opinions
  - Opinion spam detection
- Beyond Sentiments

Aspect-based Sentiment Analysis

Review 1

"I bought an iPhone a few days ago. It is such a nice phone,…"

Review 2

" The touch screen is really cool. The voice quality is clear too. It is much better than my old Blackberry,…"

Review 3

" 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, …"

….

Feature Based Summary of iPhone:

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.

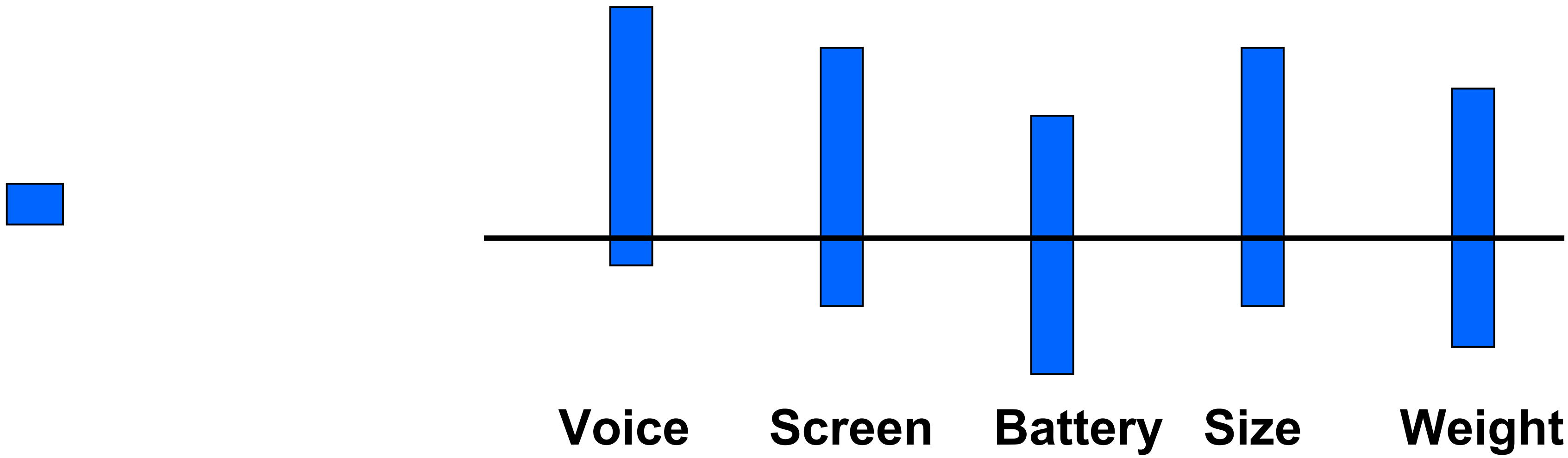
…

Feature2: **voice quality**

…

Note: We omit opinion holders

Aspect-based Sentiment Analysis



Extracting Features/Aspect

A frequency-based approach

Supervised learning

Double propagation E.g., "The **rooms** are **spacious**"

Rules from grammar dependency

Topic models(document generative model)

Opinion Spam Detection

finding unexpected rules and rule groups

Clustering Product Features for Opinion Mining

Zhongwu Zhai, Bing Liu, Hua Xu, Peifa Jia

WSDM'11, February 9–12, 2011

Introduction

In sentiment analysis of product reviews, one important problem is to **produce a summary of opinions based on product features/aspects**.

However, for the **same feature, people can express it with many different words or phrases**.

The picture quality is great.
The image looks vivid.

Roadmap

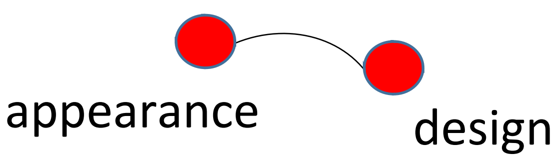
- Sentiment Analysis and Opinion Mining
  - Document sentiment classification
  - Sentence subjectivity & sentiment classification
  - Aspect-based sentiment analysis
  - Mining comparative opinions
  - Opinion spam detection
- Beyond Sentiments

Related Work

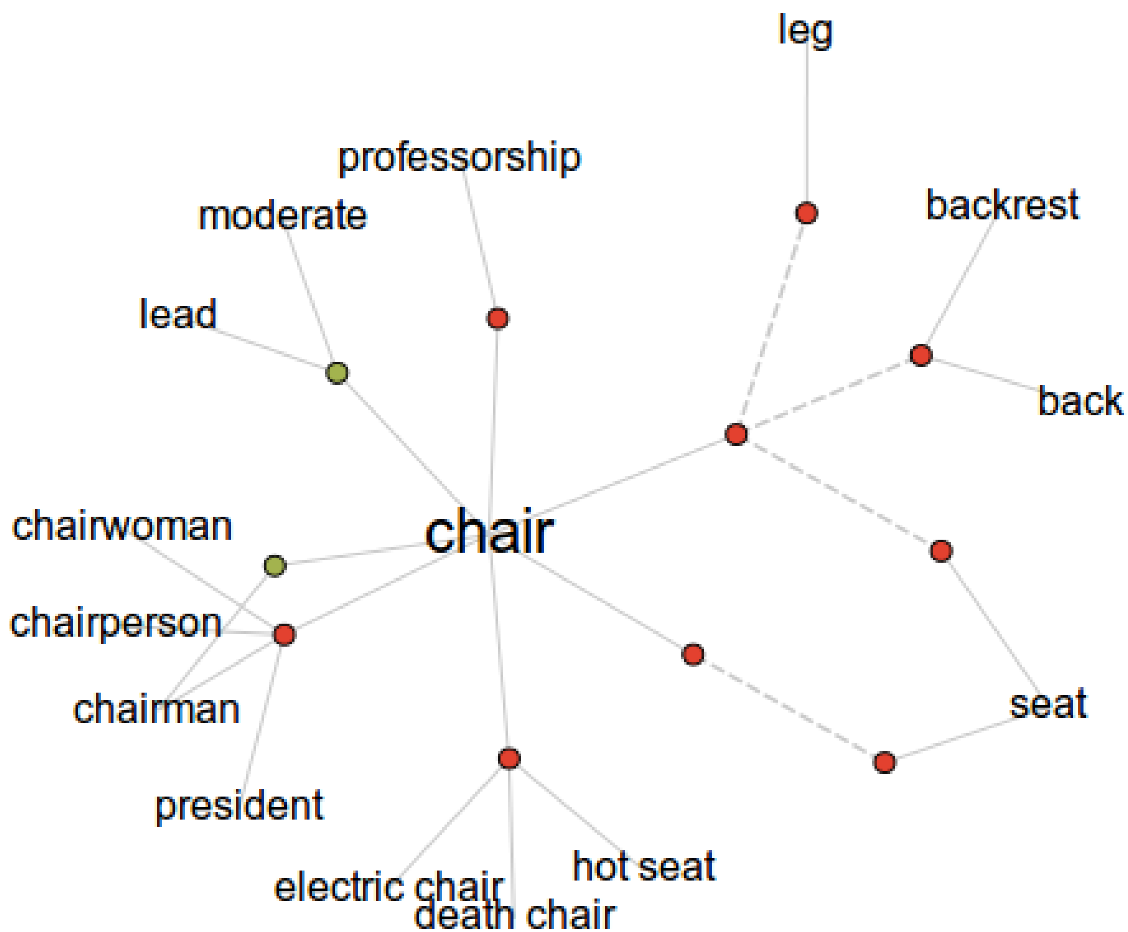
similarity measure

pre-existing knowledge resources (e.g., WordNet, and semantic networks)

- Carenini G, Ng R, and Zwart E. Extracting knowledge from evaluative text. ICKC. 2005 **(lexical similarity)**
- Liu B, Hu M, and Cheng J. Opinion Observer: Analyzing and Comparing Opinions on the Web. WWW. 2005



WordNet


leg
professorship
backrest
moderate
lead
back
chairwoman
chair
chairperson
chairman
seat
president
hot seat
electric chair
death chair

Related Work

distributional properties of words in corpora

- Bollegala D, Matsuo Y, and Ishizuka M. Measuring semantic similarity between words using web search engines. WWW. 2007
- Pedersen T. Information Content Measures of Semantic Similarity Perform Better Without Sense-Tagged Text. NAACL HLT. 2010

Related Work

topic modeling

- Branavan S R K, Chen H, Eisenstein J, and Barzilay R. Learning document-level semantic properties from free-text annotations. ACL. 2008
- Andrzejewski D, Zhu X, and Craven M. Incorporating domain knowledge into topic modeling via Dirichlet forest priors. ICML. 2009

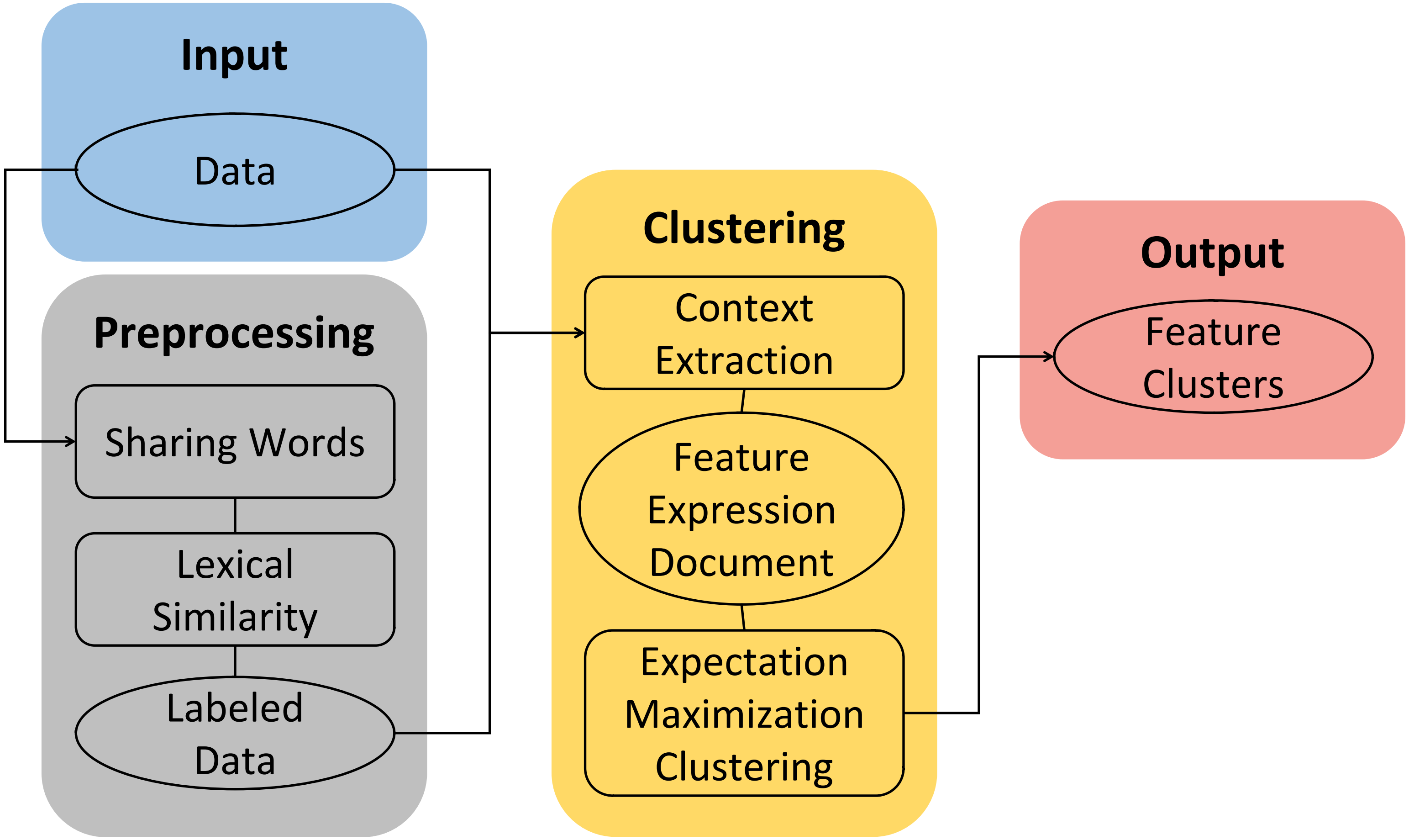
Contributions of Paper

The problem solved in this paper is **semi-supervised learning** task but without asking the user to label any training examples.

They propose two **soft constraints** to help label some examples and one piece of pre-existing natural language knowledge to extract more discriminative distributional context for the augmented EM.

An EM algorithm based on naïve Bayesian classification is adapted to solve the problem, which **allows EM to re-assign classes** of the labeled examples to different classes.

Architecture



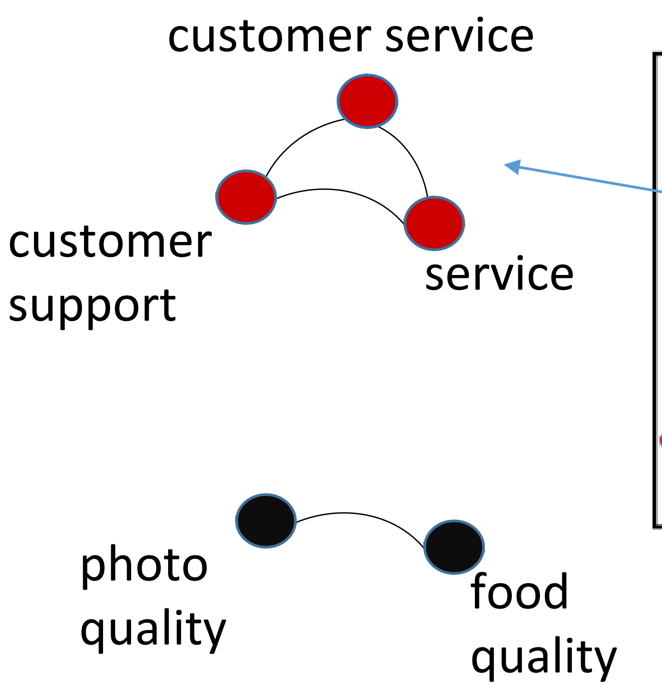
Algorithm

Input: a set of reviews **R**, and a set of discovered feature expressions **F** from **R**

1. **Generating Labeled Data L**
    1) Connect feature expressions using **sharing words**
    2) Merge components using **lexical similarity**
    3) Select the leader components as labeled data
2. **Semi-Supervised Learning using EM**

Generating Labeled Data L - Step 1

sharing words: feature expressions(red node) sharing some words are likely to belong to the same group or cluster.



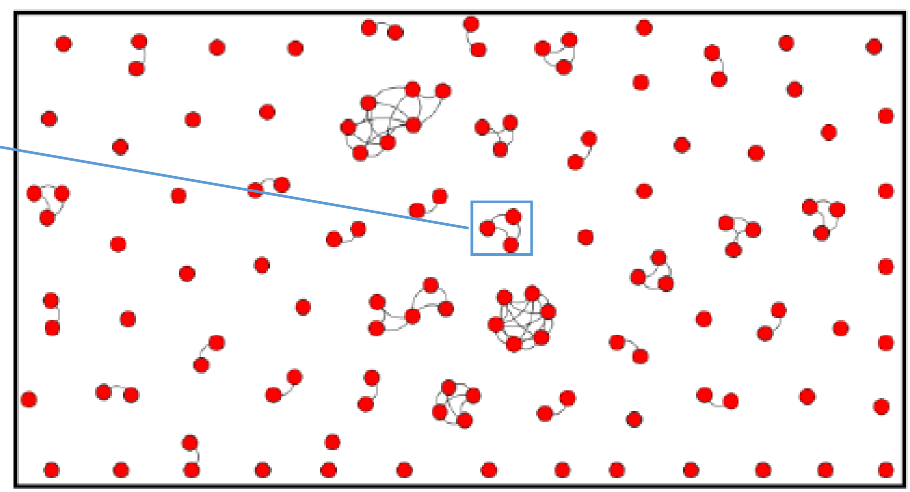


Figure 1. The G_{sc} of the Insurance data set

Generating Labeled Data L - Step 2

Input: components of G_{sc} $\{c_1, c_2, \ldots, c_n\}$;
number of merges K;
Output: merged components $\{C_1, C_2, \ldots, C_p\}$

```
//Calculate pairwise similarities of {c_1, c_2, ..., c_n}
for c_i in {c_1, c_2, ..., c_n}:
    for c_j in {c_{i+1}, c_{i+2}, ..., c_n}:
        sim(c_i, c_j) = Avg_{v_r∈c_i, v_t∈c_j}(PhraseSim(v_r, v_t))
Sort pair(i, j) as SortedPairs by sim(c_i, c_j) in descending
order, where i≠j, i∈{1,2,...n}, j∈{1,2,...n}
for pair(i, j) in {top K SortedPairs}:
    merge c_i and c_j in G_sc
Output components in G_sc as {C_1, C_2, ..., C_p}

//Subfunction for calculating similarity between phrases
PhraseSim(prs1, prs2):
```

erging components by lexical similarity Figure 2. M

calculate components similarities

sort and merge top k pairs

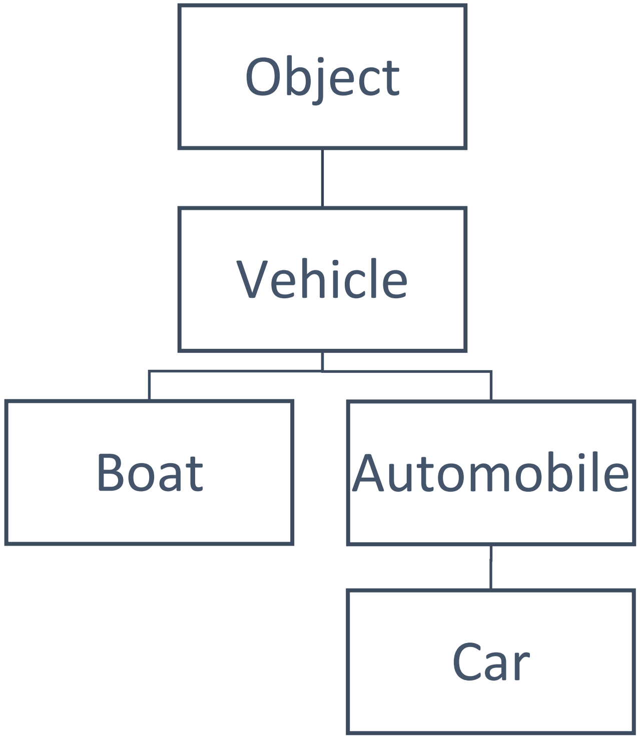
$$Res(w_1, w_2) = IC\big(LCS(w_1, w_2)\big) \quad (1)$$

$$IC(w) = -log\Pr(w) \quad (2)$$

$$Lin(w_1, w_2) = \frac{2 \times Res(w_1, w_2)}{IC(w_1) + IC(w_2)} \quad (3)$$

$$Jcn(w_1, w_2) = \frac{1}{IC(w_1) + IC(w_2) - 2 \times Res(w_1, w_2)} \quad (4)$$

Generating Labeled Data L - Step 2



Least Common Subsumer(LCS) of two concepts A and B is **the most specific concept which is an ancestor of both A and B**, where the concept tree is defined by the is-a relation.

Generating Labeled Data L - Step 2

$$Res(w_1, w_2) = IC(LCS(w_1, w_2)) \quad (1)$$

$$\boxed{IC(w) = -log\Pr(w) \quad (2)}$$

$$Lin(w_1, w_2) = \frac{2 \times Res(w_1, w_2)}{IC(w_1) + IC(w_2)} \quad (3)$$

$$Jcn(w_1, w_2) = \frac{1}{IC(w_1) + IC(w_2) - 2 \times Res(w_1, w_2)} \quad (4)$$

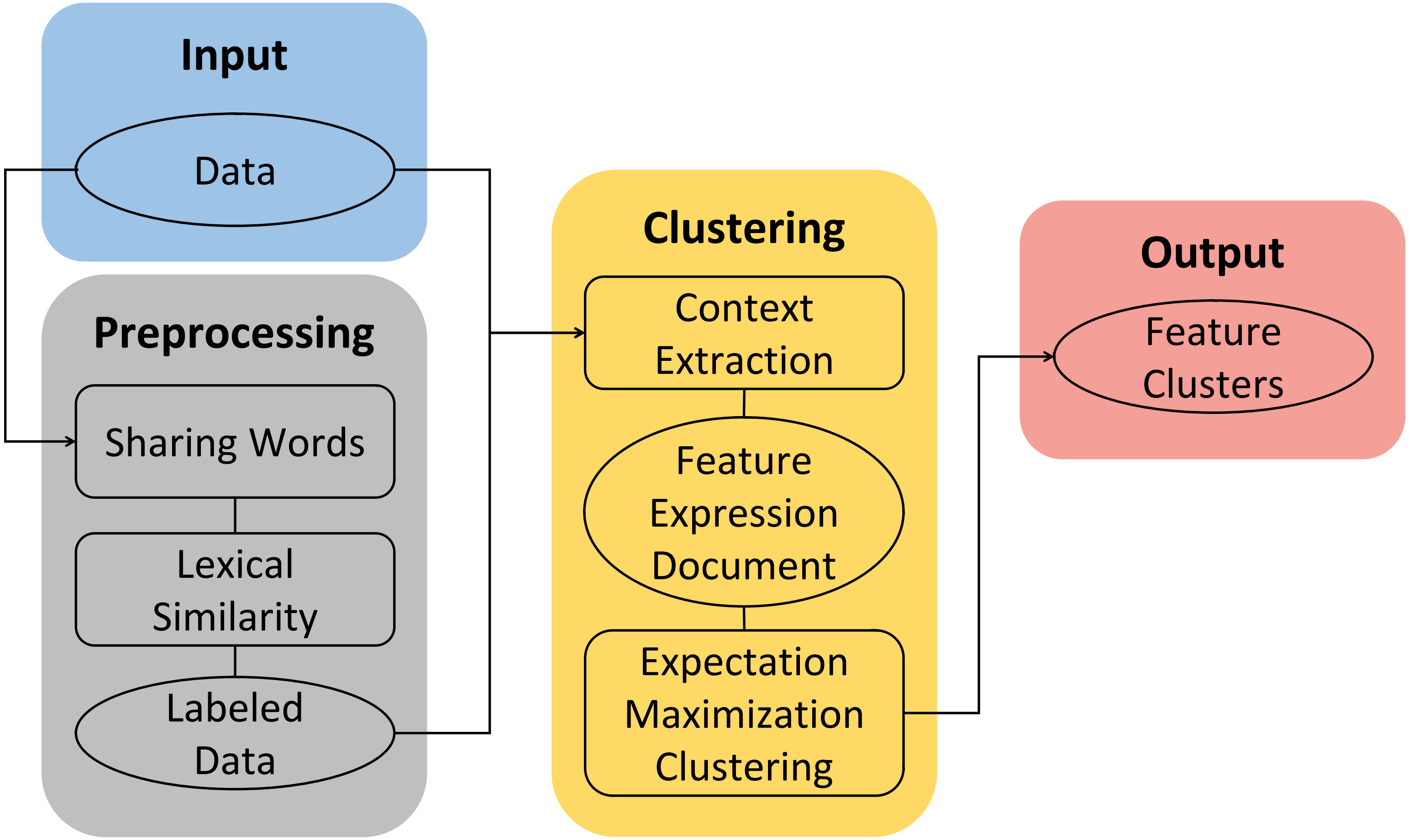
Pr(w) is the probability of the concept word **w**, based on the observed frequency counts in the WordNet corpus

two formula which calculate similarity between two words

Generating Labeled Data L - Step 3

This step **selects k leader components from the all components** to form the labeled data with **k** classes or clusters.

Architecture



Content Extraction - Example

“The **picture quality** is great, the **battery life** is also long, but the **zoom** is not good.”

Content Extraction

Labeled Example **L**
Unlabeled Example **U**

For example, a feature expression from **L** (or **U**) is

$\mathbf{v_i}$ =“screen”

$\mathbf{s_{i1}}$ = “The LCD **screen** gives clear picture”.

$\mathbf{s_{i2}}$ = “The **screen** is too small”.

$\mathbf{D_i}$ = {<LCD, screen, gives, clear>, <screen, small>}

Semi-Supervised Learning using EM

Input: Labeled examples L
Unlabeled examples U

```
1  Learn an initial naïve Bayesian classifier f_0 using L and
   Equations 5 and 6;
2  repeat
3     // E-Step
4     for each example d_i in U ∪ L :
5         Using the current classifier f_x to compute P(c_j|d_i) using
          Equation 7.
6     end
7     // M-Step
8     Learn a new naïve Bayesian classifier f_x from L and U by
      computing P(w_t|c_j) and P(c_j) using Equations 5 and 6.
9  until the classifier parameters stabilize
```

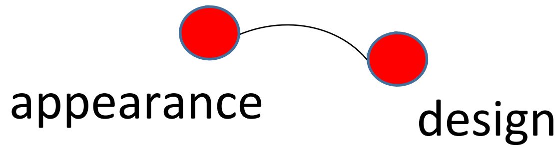
Output: the classifier f_x from the last iteration.

Figure 3. The augmented EM algorithm

Semi-Supervised Learning using EM

Distributional information is critical for finding domain synonyms because it gives the domain context.

The approach is able to **explore both domain independent** **(pre-existing knowledge)** **and domain dependent** **(distributional information)**.



Evaluation – Dataset

Data Sets: Home theater(H), Insurance (I), Mattress (M), Car (C) and Vacuum (V).

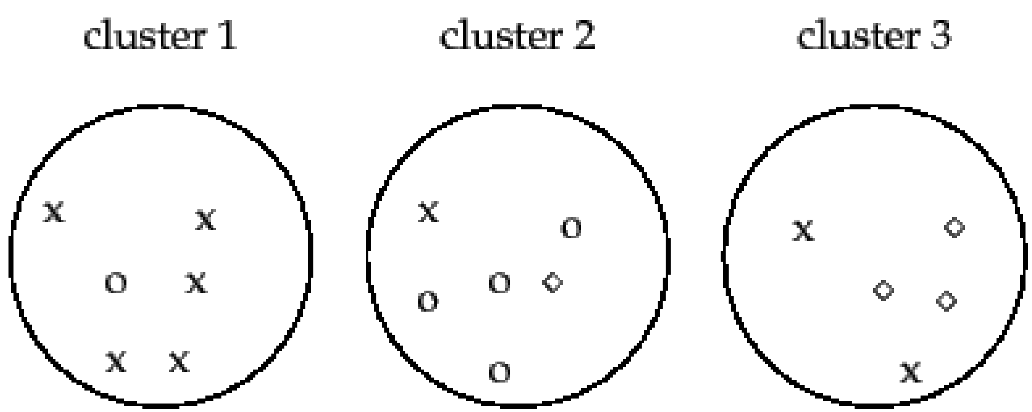
Table 1. Data sets and gold standards

	H	I	M	C	V
#Sentences	6355	12446	12107	9731	8785
#Reviews	587	2802	933	1486	551
#Expressions	237	148	333	317	266
#Groups	15	8	15	16	28

Evaluation – Evaluation Measures

$$entropy(DS_i) = -\sum_{j=1}^{k} \mathrm{P}_i(g_j) log_2 \mathrm{P}_i(g_j)$$

$$purity(DS_i) = \max_j \mathrm{P}_i(g_j)$$



► **Figure 16.1** Purity as an external evaluation criterion for cluster quality. Majority class and number of members of the majority class for the three clusters are: x, 5 (cluster 1); o, 4 (cluster 2); and ⋄, 3 (cluster 3). Purity is $(1/17) \times (5 + 4 + 3) \approx 0.71$.

Evaluation – Methods

Table 2. Experimental results on 5 data sets, i.e., H, I, M, C, and V.

Method	Entropy						Purity					
	H	I	M	C	V	avg	H	I	M	C	V	avg
Kmeans(TF)	**2.45**	2.32	**2.61**	2.67	**2.15**	**2.44**	**0.35**	0.31	**0.38**	0.32	**0.41**	**0.35**
Kmeans(PMI)	2.87	2.37	2.80	2.99	2.66	2.74	0.26	0.35	0.29	0.24	0.27	0.28
LDA	2.63	2.32	2.75	2.63	2.43	2.55	0.31	0.35	0.31	0.32	0.35	0.33
mLSA	2.62	**2.31**	2.74	**2.46**	2.38	2.50	0.32	**0.35**	0.31	**0.38**	0.37	0.35
Newman(χ^2)	2.88	2.48	2.75	2.89	2.54	2.71	0.27	0.32	0.32	0.26	0.34	0.30
Newman(PMI)	3.06	2.42	2.98	3.31	3.64	3.08	0.26	0.30	0.26	0.21	0.23	0.25
CHC	**2.73**	**2.20**	**2.74**	**2.82**	**2.32**	**2.56**	**0.28**	**0.41**	**0.28**	**0.30**	**0.37**	**0.33**
SHC	3.35	2.55	3.15	3.41	3.76	3.24	0.25	0.34	0.23	0.21	0.23	0.25
L-Rand	2.18	2.19	2.52	2.58	1.84	2.26	0.47	0.41	0.42	0.38	0.52	0.44
L-Kmeans(TF)	1.90	2.04	2.24	2.18	1.56	1.98	0.52	0.43	0.46	0.47	0.57	0.49
L-Kmeans'(TF)	1.95	1.91	2.39	2.29	1.85	2.08	0.51	0.46	0.44	0.45	0.50	0.47
L-LDA	2.12	2.11	2.24	2.25	1.70	2.08	0.46	0.43	0.48	0.45	0.55	0.47
DF-LDA	2.19	1.91	2.11	2.14	1.64	2.00	0.41	0.49	0.46	0.47	0.50	0.47
L-EM	**1.89**	**1.59**	**2.14**	**2.04**	**1.58**	**1.84**	**0.55**	**0.59**	**0.51**	**0.53**	**0.59**	**0.55**

Evaluation – Number of Merges (k)

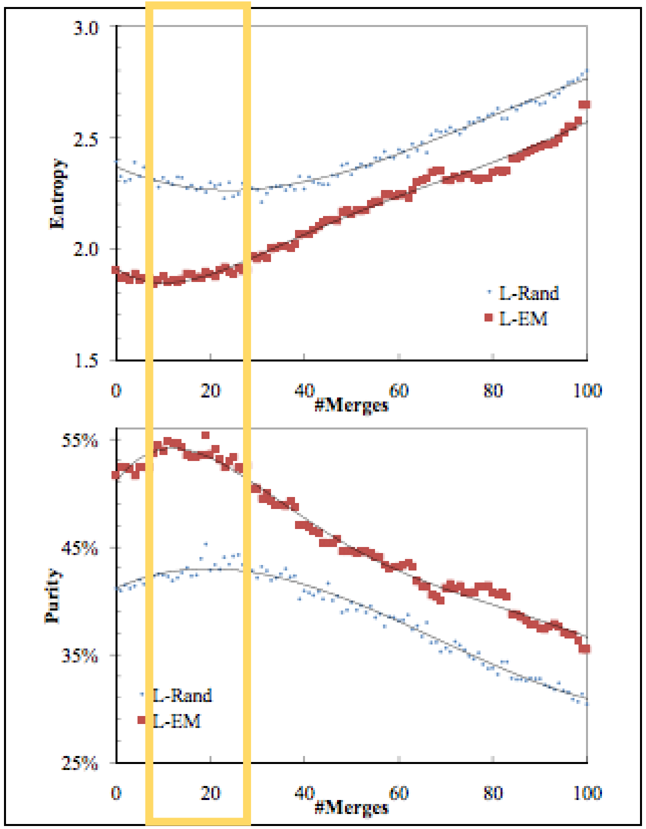


Figure 5. The influence of the number of merges to the proposed algorithm *L-EM*

Conclusions

The problem solved in this paper is **semi-supervised learning** task but without asking the user to label any training examples.

An EM algorithm based on naïve Bayesian classification is adapted to solve the problem, which **allows EM to re-assign classes** of the labeled examples to different classes.

They propose two **soft constraints** to help label some examples and one piece of pre-existing natural language knowledge to extract more discriminative distributional context for the augmented EM.

Comments - pros

Without asking the user to label

Comments

There are some domain constraint when doing opinion mining.

Not only features extract, when doing positive or negative opinion classify. We can do with semi-supervised learning.

Disappointed

2 out of 5, reviewed on Mar 28, 2013

The iPhone 6 Plus can also be an infuriating device which polarizes opinion. Apple may be widely held up as a master of modern industrial design, but for me it has got the iPhone 6 Plus completely wrong. The biggest issue is ergonomics. The iPhone 6 is a **flat**, **big** phone that doesn't feel particularly good in-hand, but just about gets away with it because it is relatively compact.

Read more