Distributional Clustering of Words for Text Categorization

Ron Bekkerman

M.Sc. Thesis

August the 8th, 2002

Text Categorization (TC)

- Learn to classify documents to one or more of predefined semantic categories
  - Supervised
  - Unsupervised
- Central issues:
  - Text Representation
  - Classifier Induction
  - Model Selection

Preliminary version of this work presented at SIGIR'01

Text Categorization (TC)

- Learn to classify documents to one or more of predefined semantic categories
  - Supervised
  - Unsupervised
- Central issues:
  - Text Representation
  - Classifier Induction
  - Model Selection

Preliminary version of this work presented at SIGIR'01

Text Representation for TC

- *Bag-of-words (BOW):* to-date, most popular representation
- Variety of other representations:
  - *N*-grams (tuples of words)
  - Sequences of characters
  - Feature clusters etc.
- Main characteristics of representations:
  - High Dimensionality
  - Statistical Sparseness
  - Level of preserving semantic relations

Text Representation for TC

- *Bag-of-words (BOW):* to-date, most popular representation
- Variety of other representations:
  - *N*-grams (tuples of words)
  - Sequences of characters
  - Feature clusters etc.
- Main characteristics of representations:
  - High Dimensionality
  - Statistical Sparseness
  - Level of preserving semantic relations

Example: BOW

d

The ceremony may assist in emphasizing the depth of such a commitment, but is of itself nothing. God knows our hearts. He knows when two have committed themselves to be one, he knows the fears and delusions we have that keep us from fully giving ourselves to another.

- A document from 20NG (soc.religion.christian)

Example: BOW

d

The ceremony may assist in emphasizing the depth of such a commitment, but is of itself nothing. God knows our hearts. He knows when two have committed themselves to be one, he knows the fears and delusions we have that keep us from fully giving ourselves to another.

- A document from 20NG (soc.religion.christian)
- Representation: vector of ~50,000 elements

Bow(d)

Example: BOW

d

The ceremony may assist in emphasizing the depth of such a commitment, but is of itself nothing. God knows our hearts. He knows when two have committed themselves to be one, he knows the fears and delusions we have that keep us from fully giving ourselves to another.

- A document from 20NG (soc.religion.christian)
- Representation: vector of ~50,000 elements
- Only 40 of which are non-zero
- No relations between words are preserved

Bow(d)

Our initial approach: employ NLP

The ceremony may assist in emphasizing the depth of such a commitment, but is of itself nothing. God knows our hearts. He knows when two have committed themselves to be one, he knows the fears and delusions we have that keep us from fully giving ourselves to another.

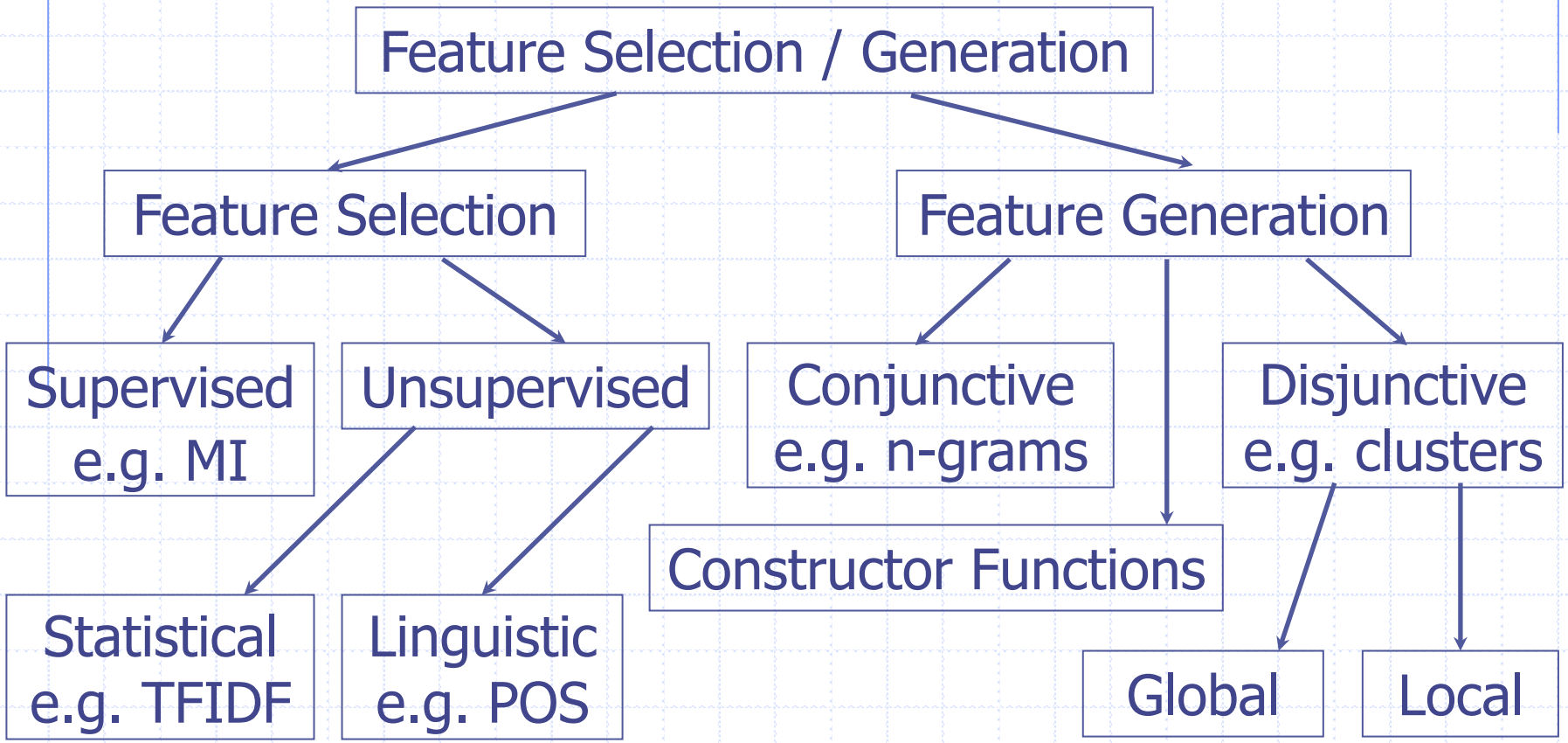
- Extract (Subject,Verb) from each sentence
- Example: 3 pairs are extracted:
  - ceremony assist
  - God knows
  - delusions keep
- Appears to capture meaning
- Yet, significant content words ignored

Content words: what are they?

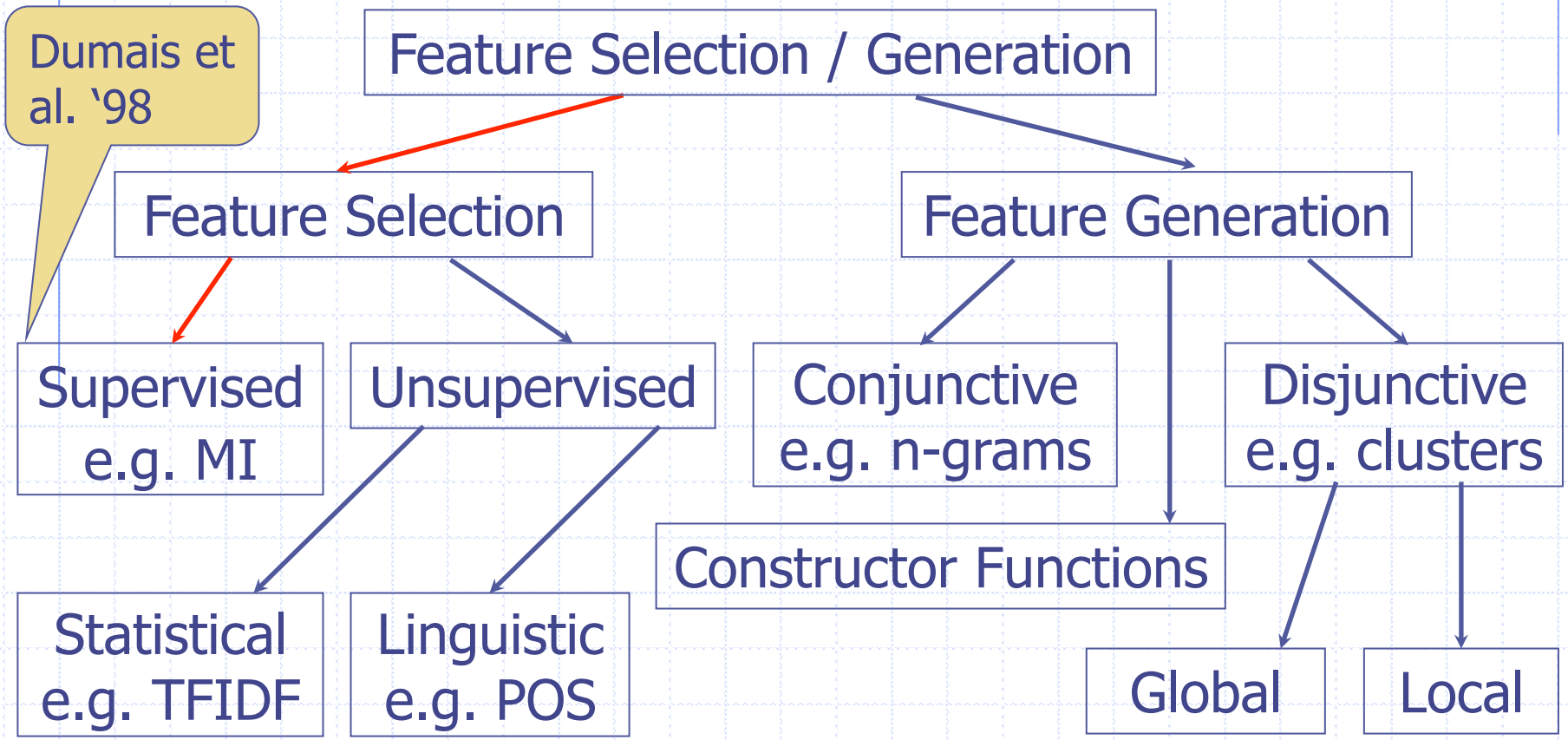
The ceremony may assist in emphasizing the depth of such a commitment, but is of itself nothing. God knows our hearts. He knows when two have committed themselves to be one, he knows the fears and delusions we have that keep us from fully giving ourselves to another.

- Syntactic roles of content words vary
- "Minor" words may compose a significant phrase
  - "fully giving ourselves to another"
- How to extract good words? Use statistics!
  - Use Feature Selection

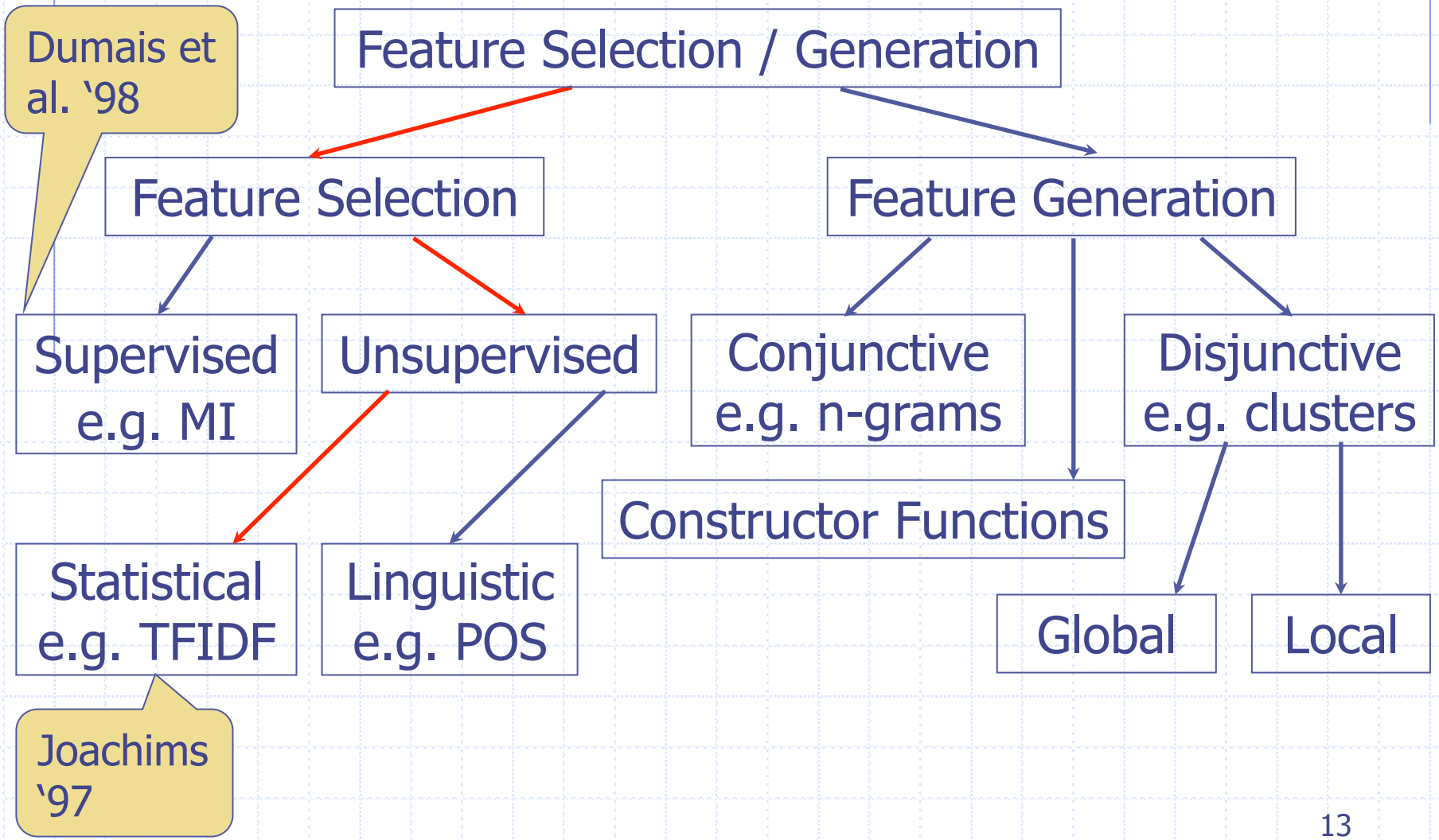
Feature Selection / Generation



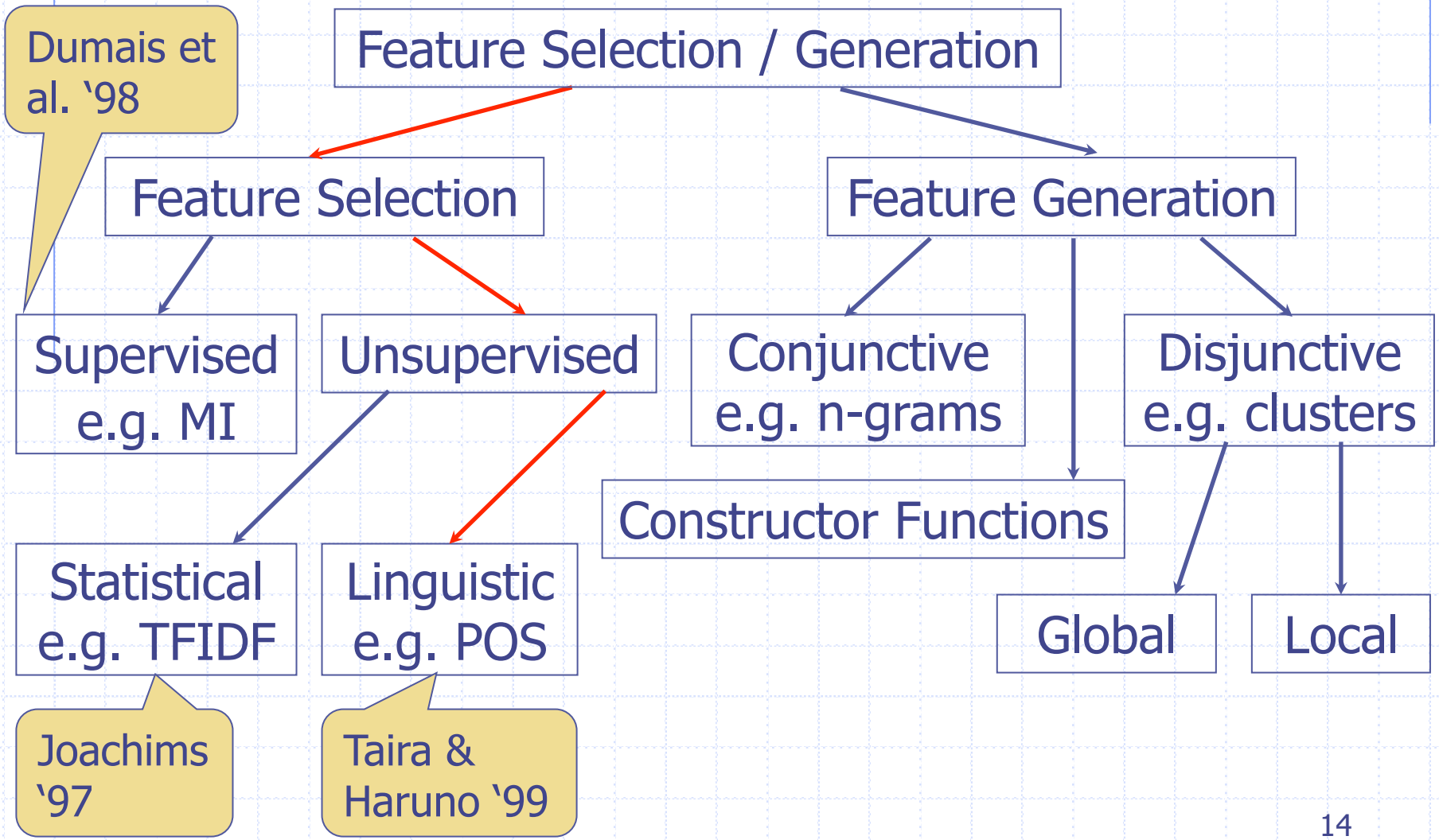
Feature Selection / Generation



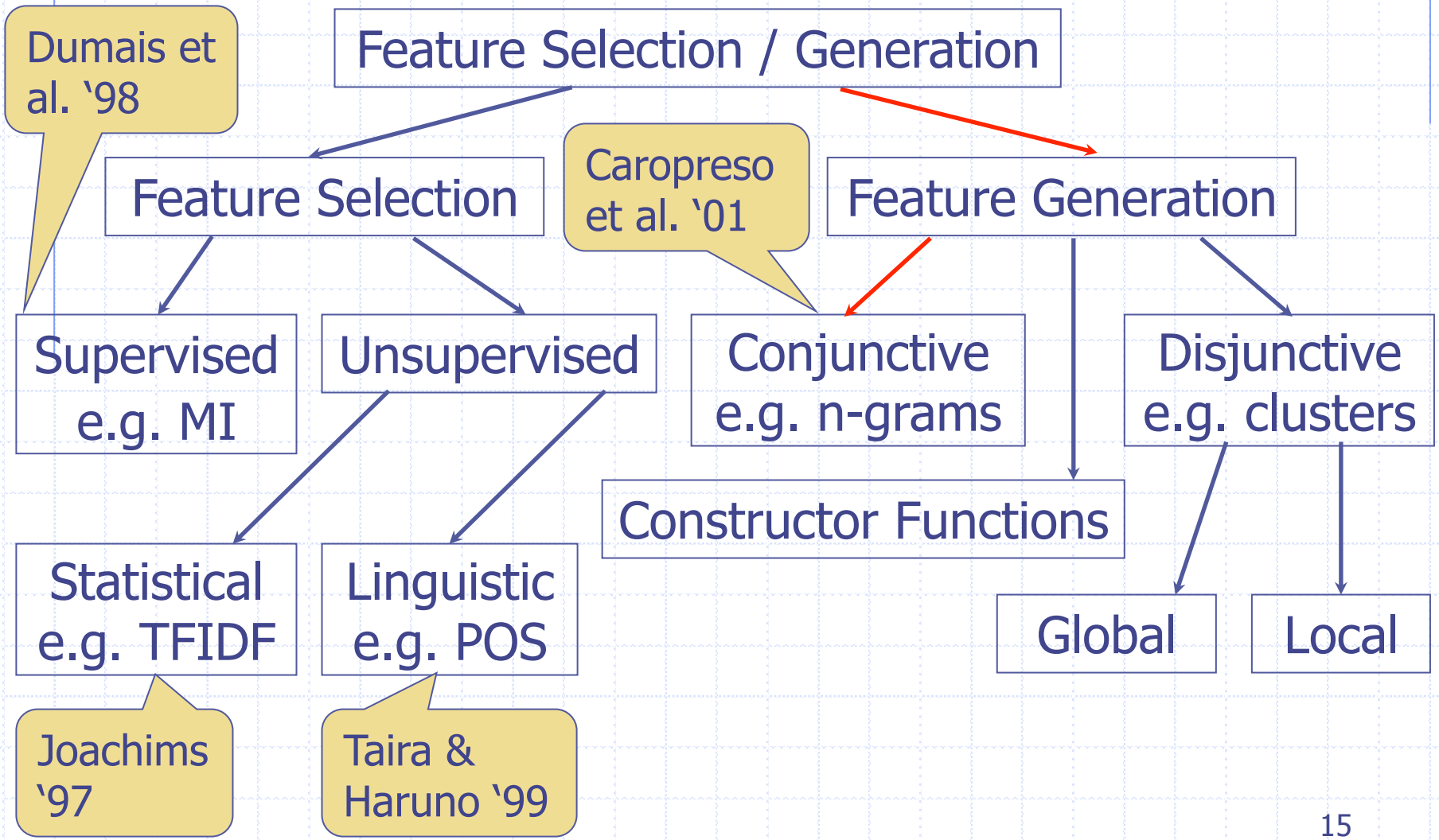
Feature Selection / Generation



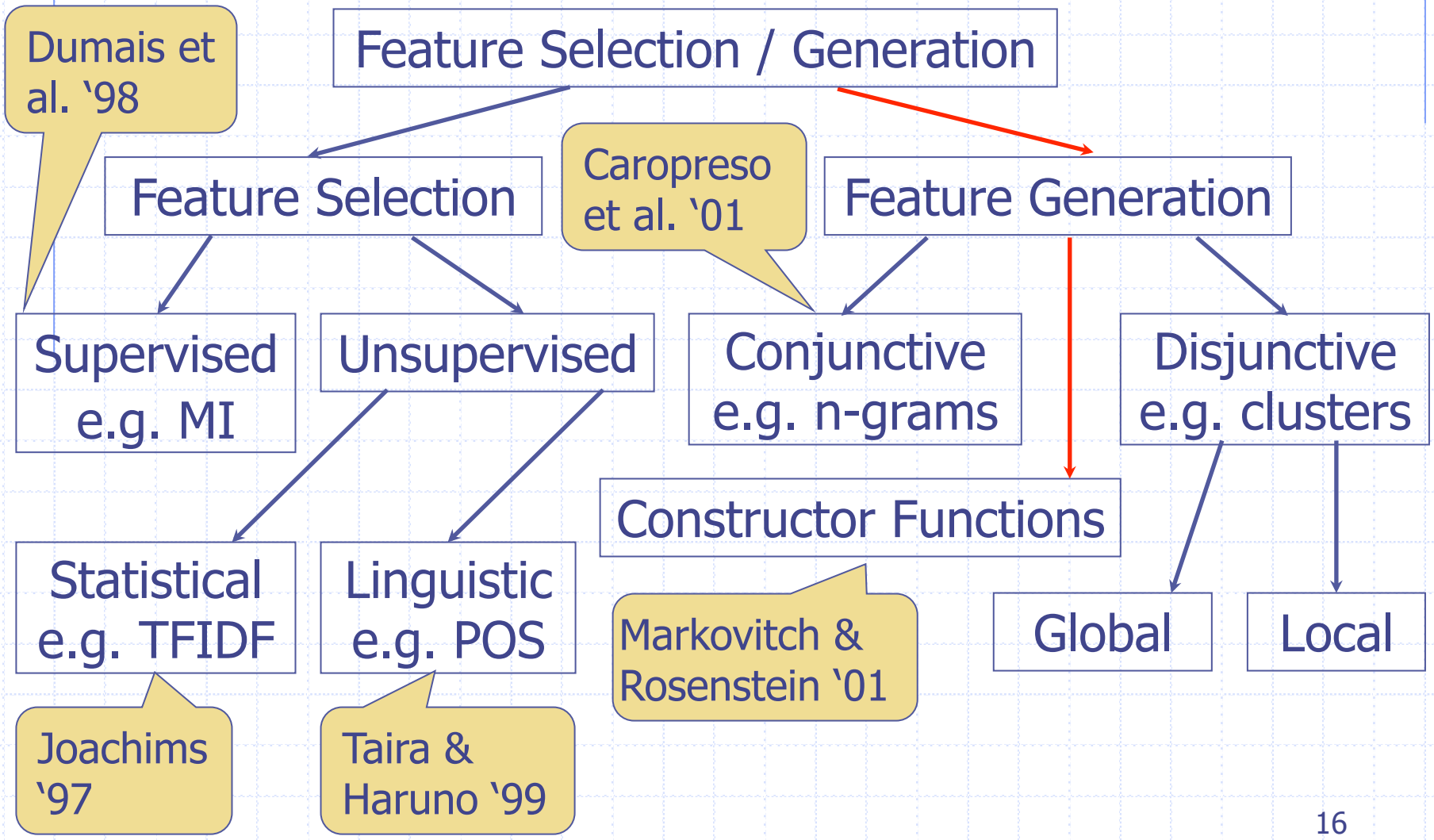
Feature Selection / Generation



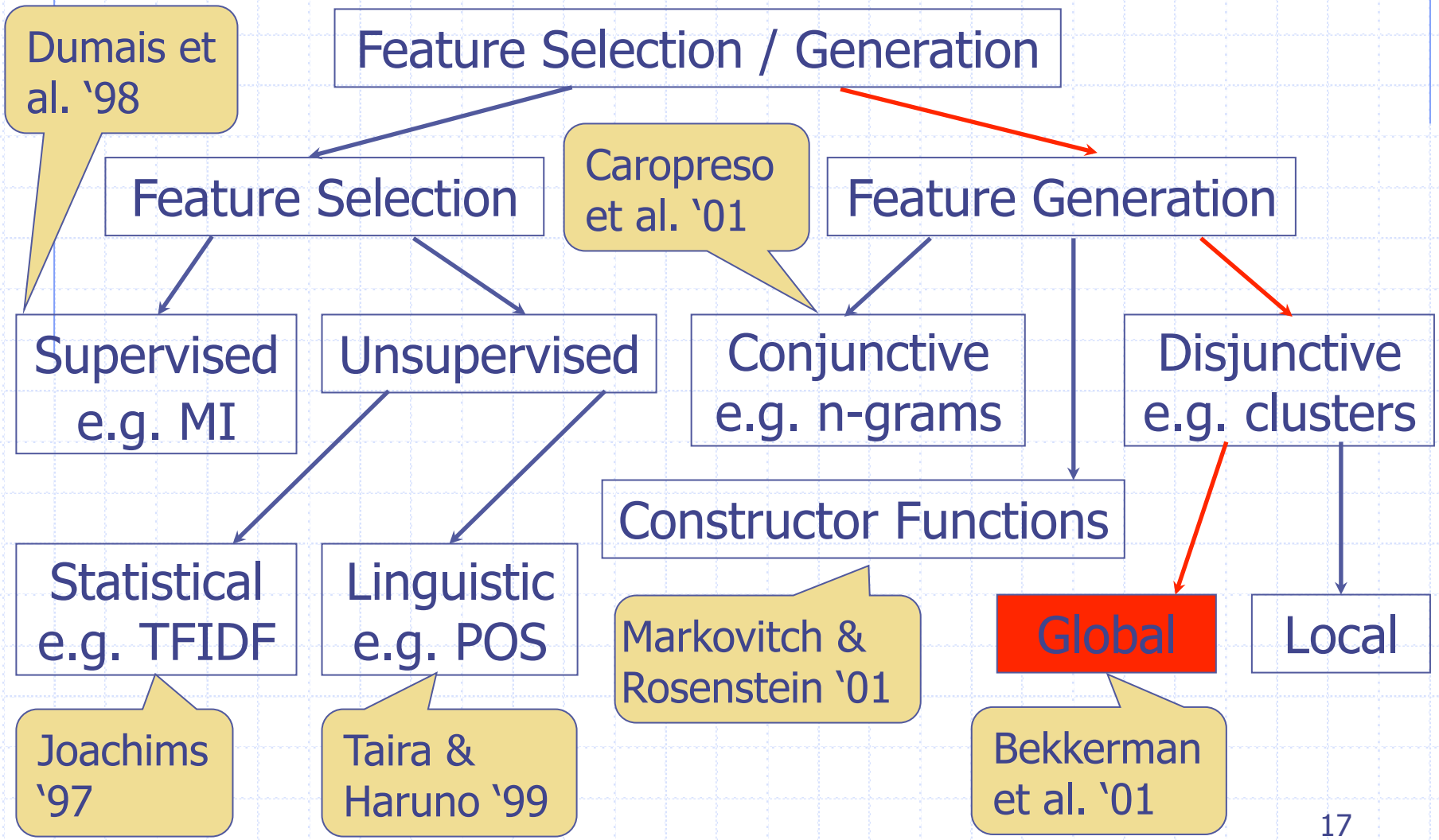
Feature Selection / Generation



Feature Selection / Generation



Feature Selection / Generation



Feature Selection using Mutual Information (MI)

The ceremony may assist in emphasizing the depth of such a commitment, but is of itself nothing. God knows our hearts. He knows when two have committed themselves to be one, he knows the fears and delusions we have that keep us from fully giving ourselves to another.

- Within 300 most discriminating words

Feature Selection using Mutual Information (MI)

The ceremony may assist in emphasizing the depth of such a commitment, but is of itself nothing. God knows our hearts. He knows when two have committed themselves to be one, he knows the fears and delusions we have that keep us from fully giving ourselves to another.

- Within 300 most discriminating words
- And within 15,000 most discriminating words

Feature Selection using Mutual Information (MI)

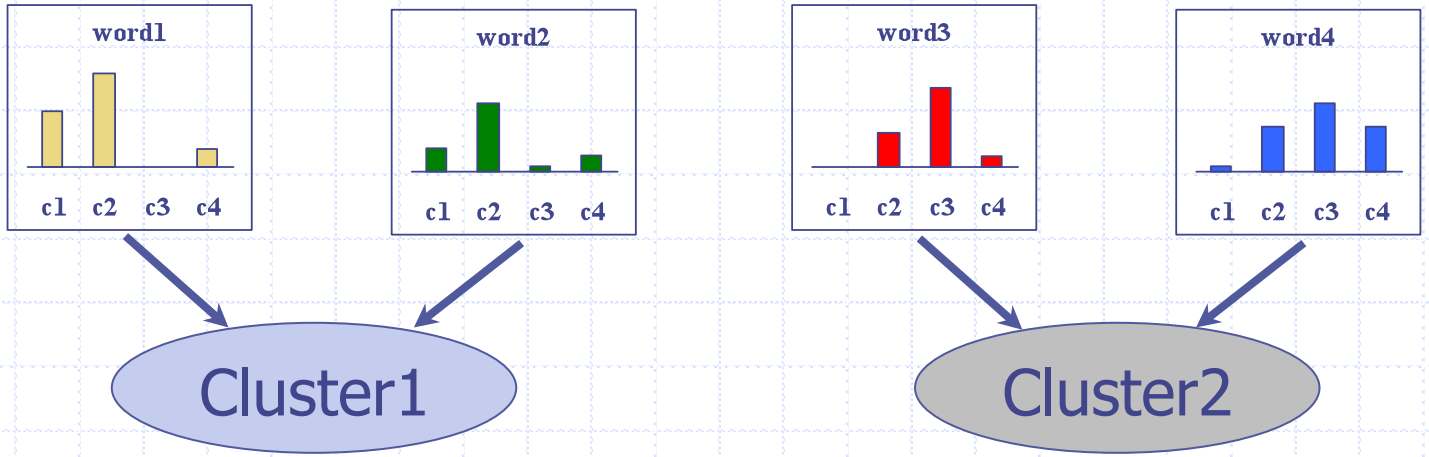
The ceremony may assist in emphasizing the depth of such a commitment, but is of itself nothing. God knows our hearts. He knows when two have committed themselves to be one, he knows the fears and delusions we have that keep us from fully giving ourselves to another.

- Within 300 most discriminating words
- And within 15,000 most discriminating words
- Insufficient words are inside:
  - "such", "another"
- Sufficient words are outside:
  - "depth", "delusions"

Our contribution

- Powerful combination of Word Distributional Clustering (via Deterministic Annealing) and SVM
- Word Distributional Clustering
  - Applied by Baker & McCallum '98
    - Simple Agglomerative Clustering + Naïve Bayes
- Support Vector Machine (SVM)
  - To-date, the best classifier from the shelf
- Our results: among the best ones ever achieved on 3 benchmarks

Word Distributional Clustering



- Solutions for the 3 problems:
  - **(High dimensionality)** The dimension is *k* (fixed)
  - **(Sparseness)** Many words mapped onto the same cluster
  - **(Semantic relations)** The set of clusters is a sort of thesaurus

Information Bottleneck (IB)

- Proposed by Tishby, Pereira & Bialek ('99)
- The idea is to construct a partition $\tilde{w}$ so that to maximize the $I(\tilde{w}, c)$ under a constraint on $I(\tilde{w}, w)$.
- The solution satisfies:

$$P(\tilde{w} \mid w) = \frac{P(\tilde{w})}{Z(\beta, w)} \exp\left[-\beta \sum_c P(c \mid w) \log \frac{P(c \mid w)}{P(c \mid \tilde{w})}\right]$$

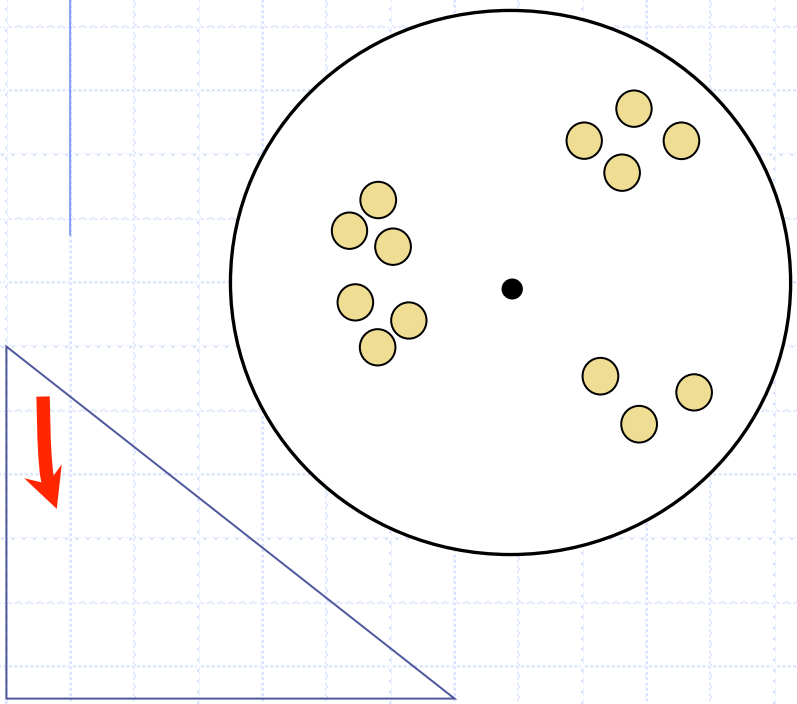
  - Z is a normalization factor, β is an annealing parameter, and $P(\tilde{w})$ $P(c \mid \tilde{w})$ are calculated using Bayes law.

IB via Deterministic Annealing

- An EM-like process should be applied
- The process is top-down
  - From 1 cluster up to *k* clusters
- 4 stages for any $1 \le i < k$:
  - Calculate $P(\tilde{w} \mid w)$ until convergence (EM)
  - Merge clusters that are too close
  - For each centroid add its "ghost"
  - Increase β (lower the temperature, as in thermodynamics)

Example: Deterministic Annealing

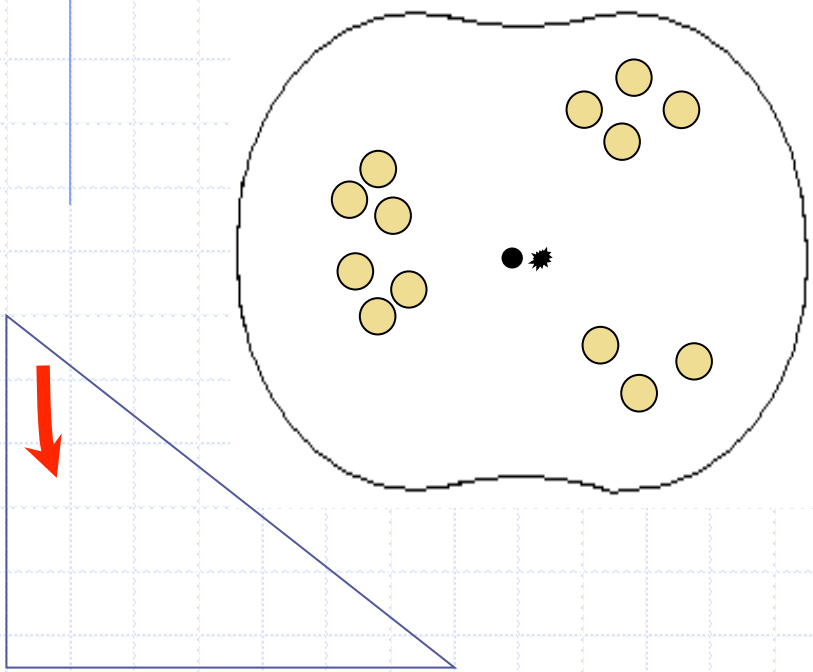
- Start with one cluster



Temperature

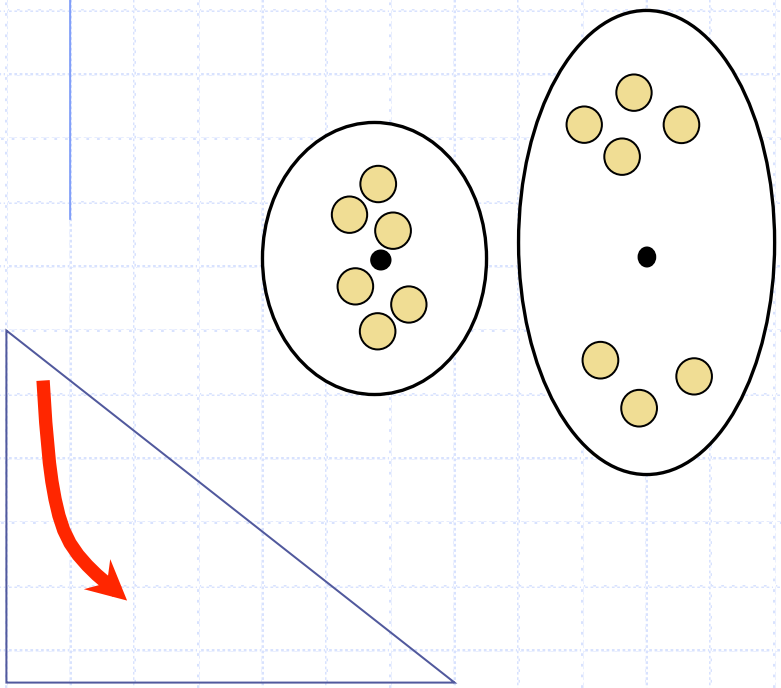
Example: Deterministic Annealing

- Start with one cluster
- Add a "ghost" centroid



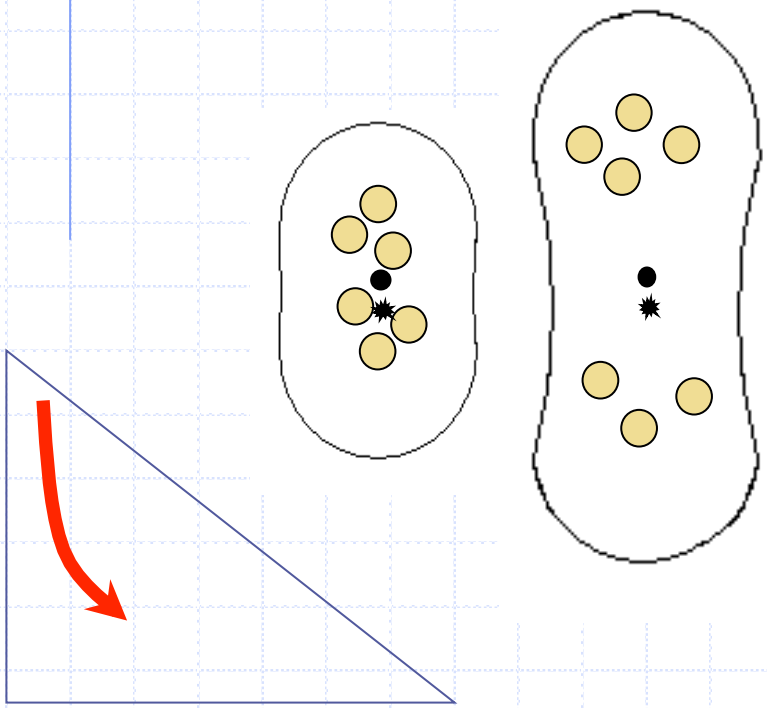
Temperature

Example: Deterministic Annealing



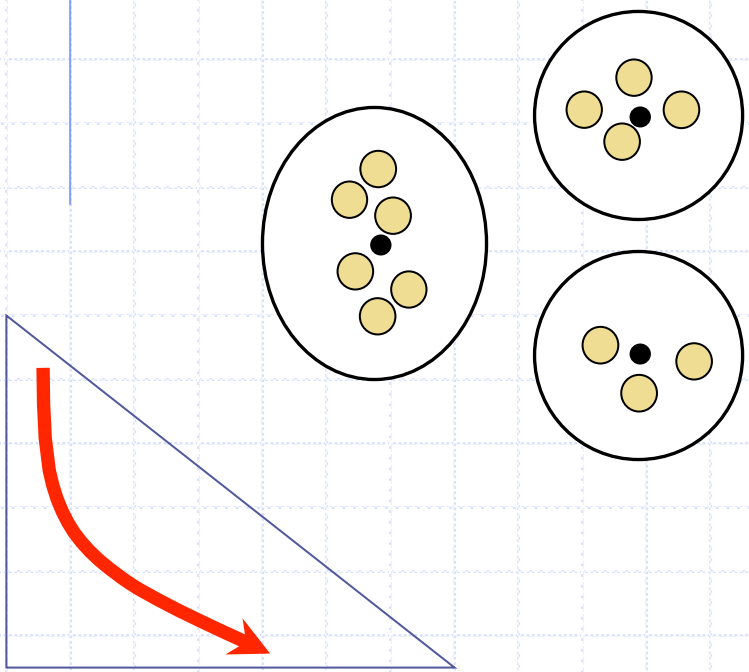
- Start with one cluster
- Add a "ghost" centroid
- Split the cluster to 2

Example: Deterministic Annealing



- Start with one cluster
- Add a "ghost" centroid
- Split the cluster to 2
- Add "ghost" centroids

Example: Deterministic Annealing

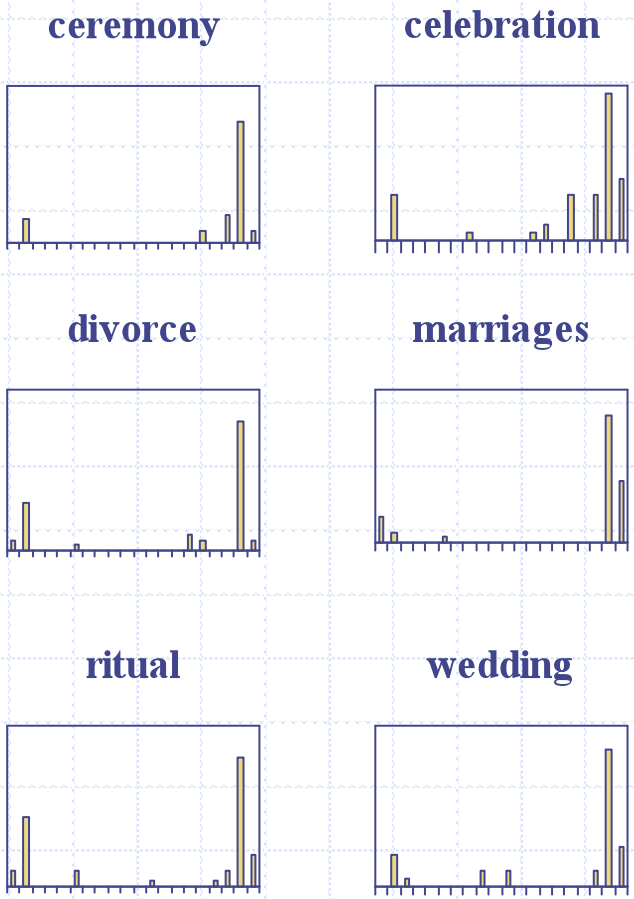


- Start with one cluster
- Add a "ghost" centroid
- Split the cluster to 2
- Add "ghost" centroids
- Split the clusters if possible

Another example: Clusters

- 1m 286 386 42bis 44m 4k 4m 61801 640x480 64k 768 8086 8500 9090 9600 accelerated accessed architecture baud bbs buffered buggy bundled card cards cd clone compatibility compatible computer computers configured connect dat dial disabling disk diskette docs fastest faxes fd formatting freeware funet hardware heine ibm install interface machine machines mag matrix megabytes memory micro mode modes mts multimedia networking optimization optimized ox pc pcs polytechnic printing proceeded processor processors resolution roms scanner scanners scanning shadows simtel simulator slower slows software svga transferring vga video wanderers
- 1835 1908 accepting accustomed acts agony ahmad appreciation arose assimilation bread brothers burial catholicism celebrated celebration ceremony charismatic condemn condemned condemns conscience consciously denounced deserts desires devastation divorce dreamed eighteenth essence father fathers feelings friendship glory grave grieve hearts heavens hebrew hindu honored humanity humble husband husbands kingdom liberating loving lust lusts majesty mankind marriages marry martyrdom materialistic missionaries moses natures obeyed orphan orthodox ourselves palms patriarchal pesach pilgrimage poetry prayed praying preach priests proclamation profess punished punishment qualities reformer refusing refutations reject rejecting rejection relationship righteous righteousness ritual rome scholarly scholars scholarship senses sentiment sisters son sons souls spiritually teaching thinkers tradition traditions tribunal truth unite vatican visions visitation wedding witness witnessing

Back to the "ceremony" example

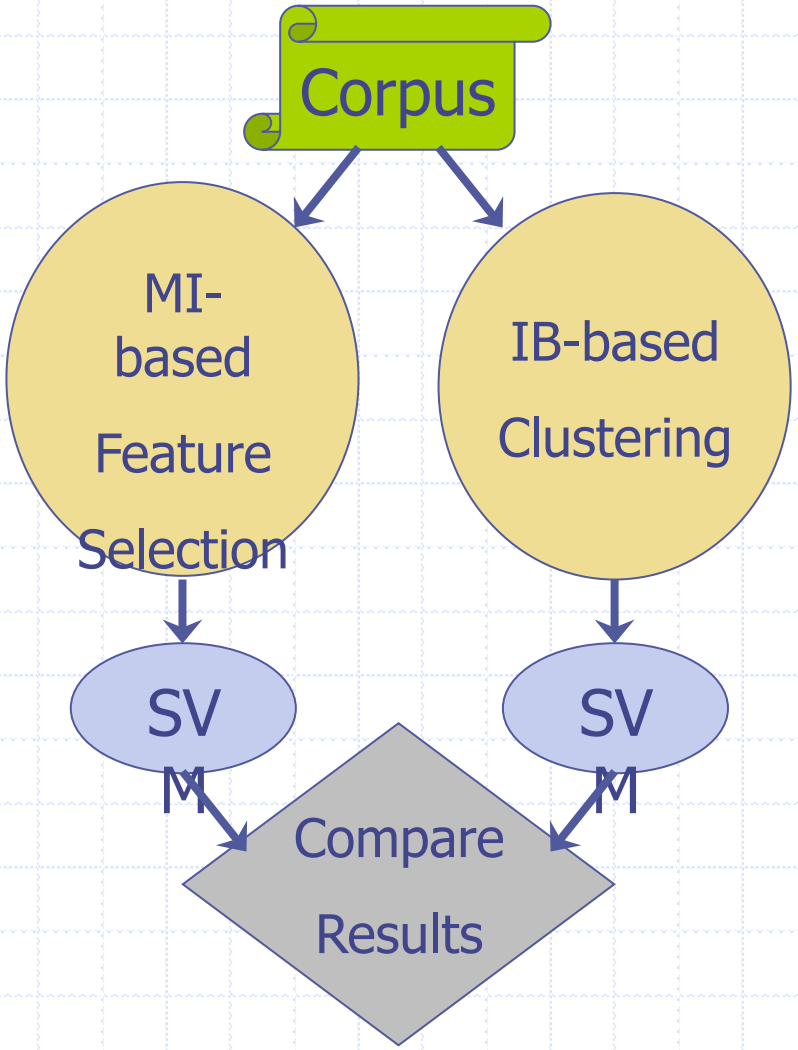


- All these words were clustered into the same cluster
- Now we know:
  - The document is about wedding!
  - Our method allows to recognize the topic
- Word Distributional Clustering is good for Text Categorization

3 Benchmark Corpora

- Reuters (ModApte Split):
  - 7063 articles in the training set, 2742 articles in the test set. 15.5% are multi-labeled
  - We consider its 10 largest categories
- 20 Newsgroups (20NG):
  - 19,997 articles, 20 categories
  - 4.5% are multi-labeled
- Web Knowledge Base (WebKB):
  - 4199 articles, 4 categories, uni-labeled

Experimental flow



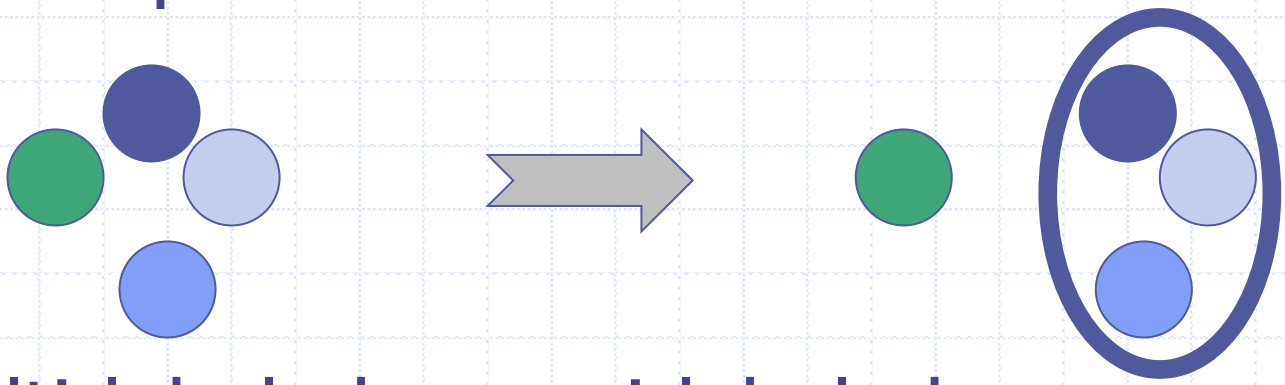
- Each document is represented as a vector of
  - Either *k* most discriminating words
  - Or *k* word clusters
- SVM is learned on a training set
- Tested on a test set

Evaluation

- 4-fold cross validation on 20NG and WebKB
  - ModApte split on Reuters
- For multi-labeled corpora:
  - Precision and Recall
    - Micro-averaged over the categories
  - Break-even point
    - For consistency with Dumais et al's work
- For uni-labeled corpora:
  - Accuracy

Issues

- Decomposition of multi-class to binary

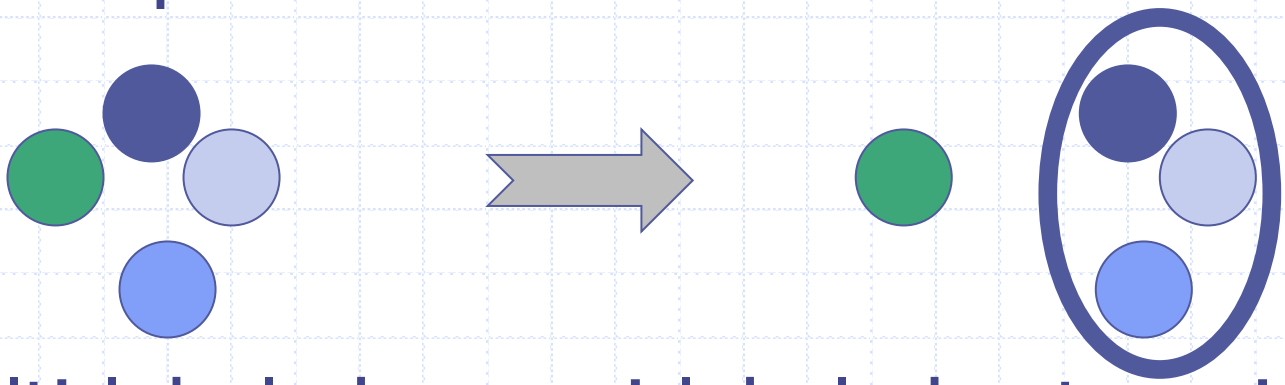


- Multi-labeled vs. uni-labeled

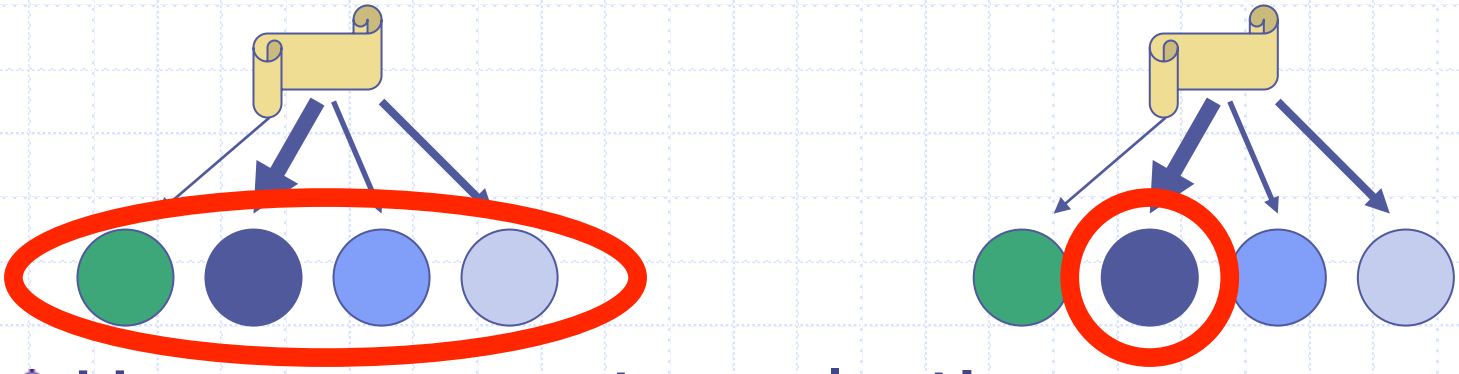
- Hyper-parameter selection

Issues

- Decomposition of multi-class to binary



- Multi-labeled vs. uni-labeled categorization



- Hyper-parameter selection

Model Selection

- Parameters were optimized on a validation set
- Sometimes we applied an "unfair" optimization
  - To emphasize empirical advantage of classifier *A* over classifier *B*, we optimized *B*'s parameters unfairly on the test set
- 4400 classifiers to build
  - Complexity reduction method used

Cl	C	J	W_{low_freq}
1	1	0.5	0
2	2	1	2
3	3	2	4
4	4	3	6
5		4	8
6		5	
7		6	
8		7	
9		8	
10		9	
11		10	
12			
13			
14			
15			
16			
17			
18			
19			
20			

Multi-labeled results

Categorizer	Reuters (BEP)	20NG (BEP)
BOW+MI k=300	92.0 (published by Dumais et al.)	77.7±0.5 (unfair) 76.5±0.4 (fair)
BOW+MI k=15000	92.0	86.3±0.5 (unfair) 85.6±0.6 (fair)
IB k=300	91.2 (fair) 92.6 (unfair)	88.6±0.3

Uni-labeled results

Categorizer	WebKB (accuracy)	20NG (accuracy)
BOW+MI k=300	92.6±0.3	85.5±0.7 (unfair) 84.7±0.7 (fair)
BOW+MI k=15000	92.4±0.5	90.9±0.2 (unfair) 90.2±0.3 (fair)
IB k=300	91.0±0.5 (unfair) 89.5±0.7 (fair)	91.3±0.4

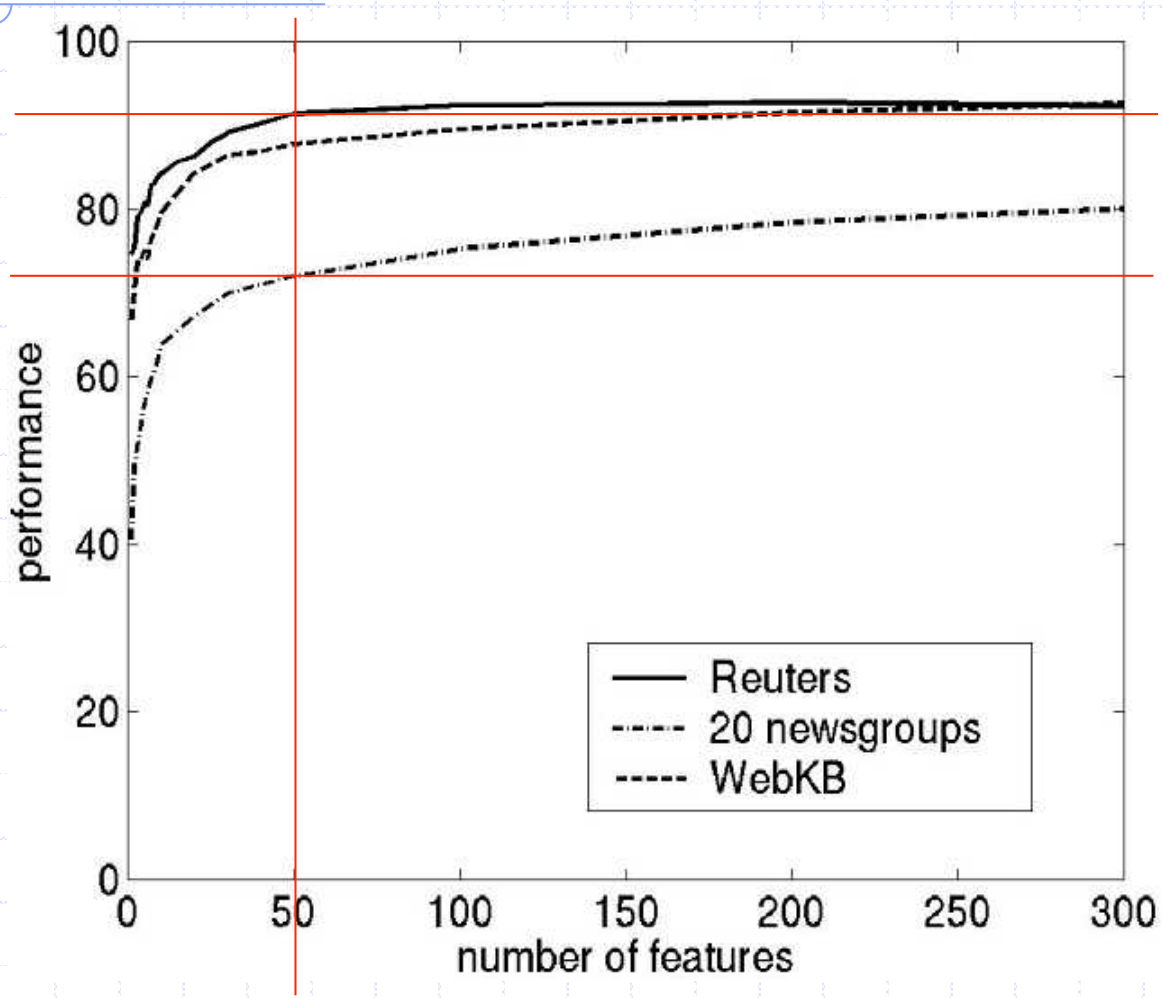
Computational Intensity

- Corpora sizes:
  - 20NG: ~40M, WebKB: ~22M, Reuters: ~11M
- Computer power:
  - Pentium III 600MHz, 2G RAM
- One run on 20NG:
  - Multi-labeled: ~2 days, uni-labeled: ~4 days
- One run on WebKB: ~1 day, Reuters: less
- Necessary runs: ~100, actually: many more
- About half a year of computer time

Discussion of the results

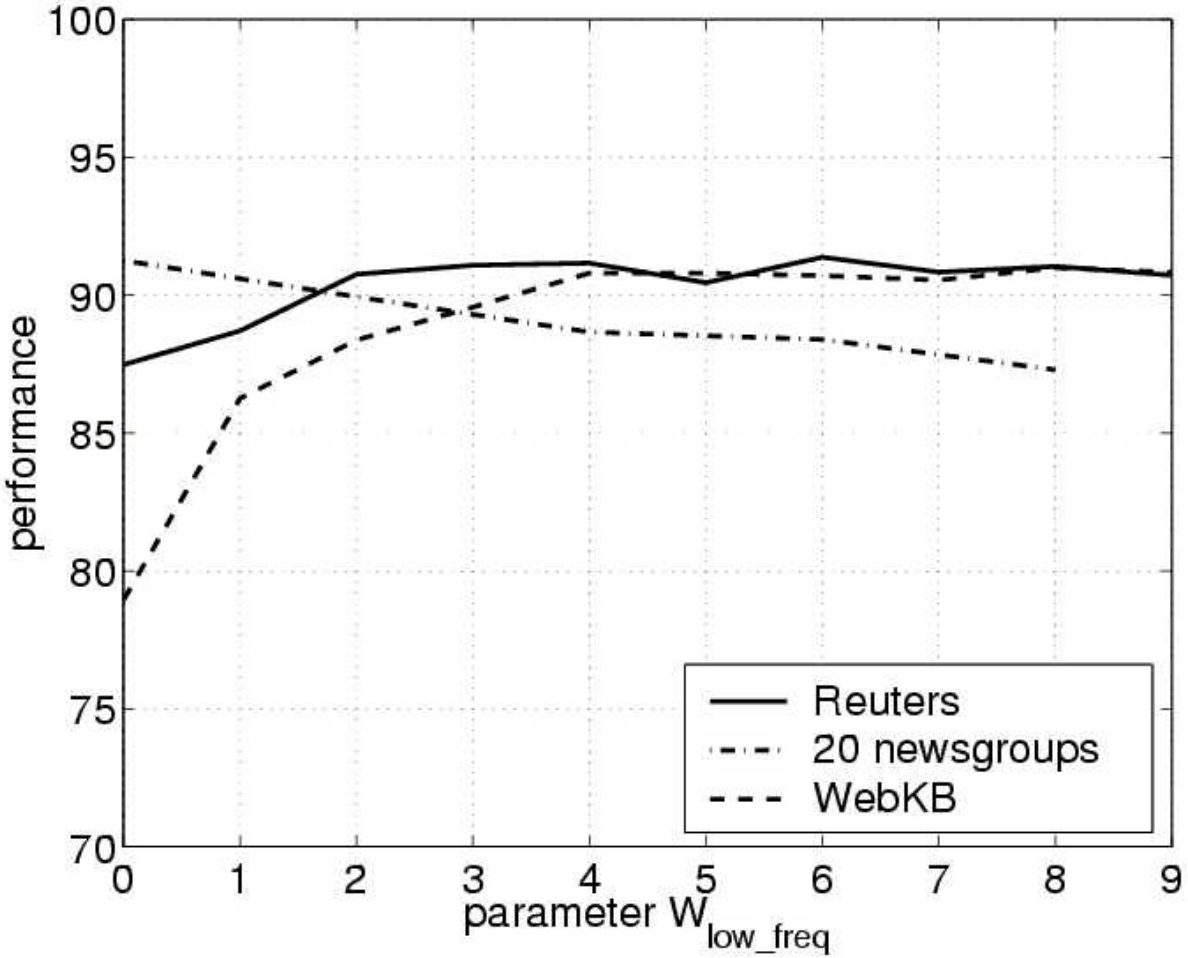
- On 20NG the IB categorizer sufficiently outperforms the BOW+MI categorizer
  - Either in categorization accuracy or in representation efficiency
- On Reuters and WebKB the IB categorizer is slightly worse
- Hypothesis: **Reuters and WebKB** and **20NG** are principally different!

BOW+MI setup: the difference



- Reuters and WebKB reach their plateau with *k≈50*
- On 20NG the result with *k=50* is ~70% while the best result is ~90%

IB setup: the difference



- Low frequent words are noise in Reuters and WebKB
- They are quite significant in 20NG

“Simple” vs. “complex” datasets

- Reuters and WebKB are “simple” corpora
  - Many documents are tables
  - Relations between words are weak
  - Keywords can be easily recognized
- 20NG is a “complex” corpus
  - Most documents are plain text
  - Texts are heterogeneous
  - Context is sufficient
- Simple Text Representation methods are satisfactory for simple corpora
  - Complex corpora require more sophisticated representation, such as word clusters

Example: "simple" datasets

- A typical Reuters document:

```
<TITLE>&lt;AIN LEASING CORP> 3RD QTR
JAN 31 LOSS</TITLE>
<DATELINE>    GREAT NECK, N.Y., March
30 -
    </DATELINE><BODY>Shr loss six cts
vs profit 22 cts
    Net loss 133,119 vs profit 496,391
    Revs 136,918 vs 737,917
    Nine mths
    Shr loss 21 cts vs profit 15 cts
    Net loss 478,991 vs profit 340,210
    Revs 324,011 vs 841,908
 Reuter
&#3;</BODY></TEXT>
```

- A typical WebKB document:

```
<html>
<body>
This page in under construction.
<p>Jimbo click below:</p>
<a href="hj1.zip"> one </a><br>
<a href="hj2.zip"> two </a><br>
<a href="hj3.zip"> three </a><br>
<a href="hj4.zip"> four</a><br>
<a href="hj5.zip"> five </a><br>
<a href="hj6.zip"> six </a><br>
<a href="hj7.zip"> seven </a><br>
</body>
</html>
```

Example: “simple” datasets

- Let us delete html tags:

```
      &lt;AIN LEASING CORP  3RD QTR
JAN 31 LOSS

              GREAT NECK, N.Y., March
30 -

                     Shr loss six cts
vs profit 22 cts

    Net loss 133,119 vs profit 496,391

    Revs 136,918 vs 737,917

    Nine mths

    Shr loss 21 cts vs profit 15 cts

    Net loss 478,991 vs profit 340,210

    Revs 324,011 vs 841,908

 Reuter

&#3;
```

```
This page in under construction.
   Jimbo click below:
        "hj1.zip"  one
        "hj2.zip"  two
        "hj3.zip"  three
        "hj4.zip"  four
        "hj5.zip"  five
        "hj6.zip"  six
        "hj7.zip"  seven
```

Example: “simple” datasets

- Let us delete non-literals:

```
        lt AIN LEASING CORP  3RD QTR
JAN 31 LOSS

              GREAT NECK  N Y   March
30

                      Shr loss six cts
vs profit 22 cts

    Net loss 133 119 vs profit 496 391

    Revs 136 918 vs 737 917

    Nine mths

    Shr loss 21 cts vs profit 15 cts

    Net loss 478 991 vs profit 340 210

    Revs 324 011 vs 841 908

 Reuter

  3
```

```
This page in under construction

   Jimbo click below

         hj1 zip   one

         hj2 zip   two

         hj3 zip   three

         hj4 zip   four

         hj5 zip   five

         hj6 zip   six

         hj7 zip   seven
```

Example: “complex” datasets

A Parable for You

"There was once our main character who blah blah blah.

"One day, a thug pointed a mean looking gun at OMC, and said, 'Do what I say, or I'm blasting you to hell.'

"OMC thought, 'If I believe this thug, and follow the instructions that will be given, I'll avoid getting blasted to hell. On the other hand, if I believe this thug, and do not follow the instructions that will be given, I'll get blasted to hell. Hmm... the more attractive choice is obvious, I'll follow the instructions.' Now, OMC found the choice obvious because everything OMC had learned about getting blasted to hell made it appear very undesirable.

"But then OMC noticed that the thug's gun wasn't a real gun. The thug's threats were make believe.

"So OMC ignored the thug and resumed blah blah blah."

Conclusion

- An effective combination of Information Bottleneck and SVM is studied
- It achieves state-of-the-art results
- On 20NG this method outperforms the simple but efficient BOW+MI categorizer
- We attempt to characterize “complex” and “simple” datasets
- “Warning” for practitioners: do not test fancy representation methods on Reuters (WebKB)

Open problems

- Given a pool of TC techniques, combine them so that the result will be as good as the best result of these techniques
  - Cross-validated Model Selection
- Use category-oriented (rather than global) clustering
- Cluster significant bigrams together with unigrams
- Tune *k* for each category

Open problem: Procedure for recognizing "simple" corpora

- Compute *N* : number of distinct words
- Apply simple MI-based feature selection
  - To extract *k* most discriminating words
- Apply 4-fold cross validation
- Learn two SVM classifiers:
  - *A* (with *k = N/2*) and *B* (with *k = N/50*)
- If *Accuracy(A)≈Accuracy(B)* the corpus is "simple" otherwise it is "complex"

Efficient Text Representation

- Representation atoms should be **words** and not strings of characters
  - Words bear Semantics !
- NLP and other corpus-independent feature extraction methods are doubtfully useful
  - Syntax → Semantics ?!
- *N*-grams may probably be useful only in combination with unigrams
- Thesauri-based representations are good
  - Level of heterogeneity decreases

A big problem of String Representation

- String Representation: the more substrings two documents have in common, the more similar the documents are
- Consider two examples:
  - "When entering the building I saw a security man who was checking bags."
  - "While coming into the house I noticed that a guard examined suitcases."
- Are the examples similar? How many substrings do they have in common?