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

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

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
Example: BOW

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

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.
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
- Supervised
  - e.g. MI
- Statistical
  - e.g. TFIDF

Feature Generation
- Unsupervised
  - Conjunctive
    - e.g. n-grams
  - Linguistic
    - e.g. POS

Constructor Functions
- Disjunctive
  - e.g. clusters
- Global
- Local
Feature Selection / Generation

Feature Selection
- Supervised: e.g. MI
- Statistical: e.g. TFIDF

Feature Generation
- Unsupervised: Conjunctive e.g. n-grams
- Linguistic: e.g. POS
- Disjunctive: e.g. clusters

Constructor Functions
- Global
- Local

Dumais et al. ‘98
Feature Selection / Generation

- Feature Selection
  - Supervised
    - e.g. MI
  - Unsupervised
    - e.g. TFIDF
- Feature Generation
  - Conjunctive
    - e.g. n-grams
  - Disjunctive
    - e.g. clusters
- Constructor Functions
  - Global
  - Local

Dumais et al. '98
Joachims '97
Feature Selection / Generation

- Feature Selection
  - Supervised
    - e.g. MI
  - Unsupervised
    - e.g. TF-IDF
- Feature Generation
  - Conjunctive
    - e.g. n-grams
  - Disjunctive
    - e.g. clusters
  - Constructor Functions
    - Global
    - Local

- Dumais et al. '98
- Joachims '97
- Taira & Haruno '99
Feature Selection / Generation

- **Supervised**
  - e.g. MI
  - Dumais et al. '98
- **Unsupervised**
  - e.g. n-grams
  - Caropreso et al. '01
- **Conjunctive**
  - e.g. n-grams
- **Disjunctive**
  - e.g. clusters
- **Statistical**
  - e.g. TFIDF
  - Statistical
  - e.g. TFIDF
  - Joachims '97
- **Linguistic**
  - e.g. POS
  - Linguistic
  - e.g. POS
  - Taira & Haruno '99
- **Constructor Functions**
  - Global
  - Local
**Feature Selection / Generation**

- **Supervised**
  - e.g. MI
  - Statistical
    - e.g. TFIDF
    - Joachims '97
  - Linguistic
    - e.g. POS
    - Taira & Haruno '99

- **Unsupervised**
  - Conjunctive
    - e.g. n-grams
    - Caropreso et al. '01
  - Disjunctive
    - e.g. clusters
    - Constructor Functions
      - Global
      - Local
      - Markovitch & Rosenstein '01

- **Taira & Haruno '99**
- **Caropreso et al. '01**
- **Dumais et al. '98**
- **Joachims '97**
Feature Selection / Generation

- **Supervised**
  - e.g. MI
  - Statistical
    - e.g. TFIDF
    - Joachims '97
  - Linguistic
    - e.g. POS
    - Taira & Haruno '99

- **Unsupervised**
  - Conjunctive
    - e.g. n-grams
    - Caropreso et al. '01
  - Disjunctive
    - e.g. clusters
    - Bekkerman et al. '01

- **Feature Selection**
  - Dumais et al. '98

- **Feature Generation**
  - Constructor Functions
    - Markovitch & Rosenstein '01
  - Global
  - Local

References:
- Taira & Haruno '99
- Bekkerman et al. '01
- Joachims '97
- Caropreso et al. '01
- Dumais et al. '98
- Markovitch & Rosenstein '01
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
    - 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:

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

- \( Z \) is a normalization factor, \( \beta \) is an annealing parameter, and \( P(\tilde{w}), P(c | \tilde{w}) \) 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 $1 \leq i < k$:
  - Calculate $P(\tilde{w} | w)$ until convergence (EM)
  - Merge clusters that are too close
  - For each centroid add its “ghost”
  - Increase $\beta$ (lower the temperature, as in thermodynamics)
Example: Deterministic Annealing

- **Start with one cluster**
Example: Deterministic Annealing

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

<table>
<thead>
<tr>
<th>$\text{Cl}$</th>
<th>$C$</th>
<th>$J$</th>
<th>$W_{\text{low}}$</th>
<th>$f_{\text{req}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.5</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>3</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>4</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>6</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>7</td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>8</td>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>9</td>
<td>9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>10</td>
<td>10</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# Multi-labeled results

<table>
<thead>
<tr>
<th>Categorizer</th>
<th>Reuters (BEP)</th>
<th>20NG (BEP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BOW+MI k=300</td>
<td>92.0 (published by Dumais et al.)</td>
<td>77.7±0.5 (unfair) 76.5±0.4 (fair)</td>
</tr>
<tr>
<td>BOW+MI k=15000</td>
<td>92.0</td>
<td>86.3±0.5 (unfair) 85.6±0.6 (fair)</td>
</tr>
<tr>
<td>IB k=300</td>
<td>91.2 (fair) 92.6 (unfair)</td>
<td>88.6±0.3</td>
</tr>
</tbody>
</table>
## Uni-labeled results

<table>
<thead>
<tr>
<th>Categorizer</th>
<th>WebKB (accuracy)</th>
<th>20NG (accuracy)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BOW+MI k=300</td>
<td>92.6±0.3</td>
<td>85.5±0.7 (unfair)</td>
</tr>
<tr>
<td>BOW+MI k=15000</td>
<td>92.4±0.5</td>
<td>90.9±0.2 (unfair)</td>
</tr>
<tr>
<td>IB k=300</td>
<td>91.0±0.5 (unfair)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>89.5±0.7 (fair)</td>
<td></td>
</tr>
</tbody>
</table>
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 $k \approx 50$
- On 20NG the result with $k=50$ is $\sim 70\%$ while the best result is $\sim 90\%$
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&gt; 3RD QTR JAN 31 LOSS</TITLE>

<DATETIME> GREAT NECK, N.Y., March 30 -
</DATETIME>
<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
</BODY></TEXT>
```

A typical WebKB document:

```
<html>
<body>
This page is 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:

<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

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:

<table>
<thead>
<tr>
<th>Lt AIN LEASING CORP 3RD QTR</th>
</tr>
</thead>
<tbody>
<tr>
<td>JAN 31 LOSS</td>
</tr>
<tr>
<td>GREAT NECK  N Y  March</td>
</tr>
<tr>
<td>30</td>
</tr>
<tr>
<td>vs profit 22 cts</td>
</tr>
<tr>
<td>Shr loss six cts</td>
</tr>
<tr>
<td>Net loss 133 119 vs profit 496 391</td>
</tr>
<tr>
<td>Revs 136 918 vs 737 917</td>
</tr>
<tr>
<td>Nine mths</td>
</tr>
<tr>
<td>Shr loss 21 cts vs profit 15 cts</td>
</tr>
<tr>
<td>Net loss 478 991 vs profit 340 210</td>
</tr>
<tr>
<td>Revs 324 011 vs 841 908</td>
</tr>
</tbody>
</table>

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 \( \text{Accuracy}(A) \approx \text{Accuracy}(B) \) 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?