Information Extraction and Summarization

March 9, 2015

Announcements

• I’m at a conference Mar 30, April 1 and will have to cancel class

• Can we have term test 2 on April 2 from 10-11:30?

Today

• Data volumes
• Strategies:
  – Visualization
  – Information extraction
  – Summarization
• Simple text processing methods
• Validation

Data

• In 2010:
  – 1.2 million petabytes of digital information (1.2 zettabytes)

• By 2020 (estimate):
  – 35 zettabytes

Grant et al., 2010

In 2010:

– 1.2 million petabytes of digital information (1.2 zettabytes)
Opportunities and Challenges

- Wealth of information
- But not in its raw form...

Data visualization

“The use of computer-generated, interactive representations of data to amplify cognition”
– [Card et al. 1999]
Information Extraction (IE)

The task of automatically extracting structured information from unstructured and/or semi-structured machine-readable documents.

-- Wikipedia

Information Extraction: Goals

1. Organize information so that it is useful to people
2. Put information in a semantically precise form that allows further inferences to be made by computer algorithms

Information Extraction: General Steps

• Find and understand limited relevant parts of texts
• Gather information from many pieces of text
• Produce a structured representation of relevant information

Example: Kylin
Summarization

- reducing text document(s) to create a summary that retains the most important points of the original(s)

Summarization: Types

- Extraction:
  - selecting a subset of existing words, phrases, or sentences in the original text to form the summary

- Abstraction:
  - build an internal semantic representation
  - use natural language generation techniques to create a summary that is closer to what a human might generate

Example: Tag Clouds

Examples: Summarizing Twitter

- Summarizing sport events via twitter [Nichols et al.]

- E.g.,
  - “Good goal for Slovenia and the USA once again starts a game terrible. Birsa gives #SVN 1-0 lead with smart shot. Howard didn’t even look like he saw that one coming.”
Some simple NLP terms

- TF-IDF
- Vector space model

TF-IDF

- Term frequency – inverse document frequency
- How “important” a word is to a document
- Often used as a weighting function in information retrieval

\[
\text{tfidf}(t, d, D) = \text{tf}(t, d) \times \text{idf}(t, D)
\]

TF Options

- raw frequency:
  \[
  \text{tf}(t, d) = f(t, d)
  \]
- Boolean "frequencies":
  \[
  \text{tf}(t, d) = 1 \text{ if } t \text{ occurs in } d \text{ and 0 otherwise}
  \]
- Logarithmic:
  \[
  \text{tf}(t, d) = 1 + \log f(t, d), \text{ or zero if } f(t, d) \text{ is zero}
  \]

IDF

\[
\text{idf}(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|}
\]
Vector Space Model

- Documents and queries represented as vectors
  - Term weights can be tf-idf
- Useful for computing document similarities

Validation

- Information retrieval metrics
  - Precision / recall
  - F
- Classifier metrics:
  - Sensitivity / specificity
- Training vs. Testing

Precision and recall

- **Precision**: % of selected items that are correct
- **Recall**: % of correct items that are selected

<table>
<thead>
<tr>
<th></th>
<th>correct</th>
<th>not correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>selected</td>
<td>tp</td>
<td>fp</td>
</tr>
<tr>
<td>not selected</td>
<td>fn</td>
<td>tn</td>
</tr>
</tbody>
</table>

Combined Measure: F

- Assesses the P/R tradeoff (weighted harmonic mean):
  \[
  F = \frac{1}{\frac{\alpha}{P} + \frac{(1-\alpha)}{R}} \frac{(\beta^2 + 1)PR}{\beta^2 P + R}
  \]
- People usually use balanced \( F_1 \) measure
  - i.e., with \( \beta = 1 \) (that is, \( \alpha = \frac{1}{2} \)):
  \[
  F = \frac{2PR}{P+R}
  \]
Classifier Metrics

- Sensitivity
  - proportion of true positives correctly identified

- Specificity
  - proportion of negatives correctly identified

Confusion Matrix

<table>
<thead>
<tr>
<th>Model/test outcome</th>
<th>True Condition</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>Positive True Positive</td>
<td>Positive predictive value (positive precision) = TP / (TP + FP)</td>
</tr>
<tr>
<td></td>
<td>False Positive</td>
<td></td>
</tr>
<tr>
<td>Negative</td>
<td>True Negative</td>
<td>Negative predictive value (negative precision) = TN / (TN + FN)</td>
</tr>
<tr>
<td></td>
<td>False Negative</td>
<td></td>
</tr>
</tbody>
</table>

- Sensitivity: $\frac{TP}{TP + FN}$
- Specificity: $\frac{TN}{FP + TN}$
- Precision: $\frac{TP + TN}{TP + FP + TN + FP}$

Sensitivity/Specificity

- For non-binary cases, these metrics are dependent on the chosen threshold

- There is a test (T4) to check if a person suffers from Hypothyroidism
- Let's suppose that we pick 5 as a threshold to declare a patient hypothyroid
- We have a sample of 125 patients, 32 with the disease and 93 healthy (Euthyroid)
- Test performance with this threshold

<table>
<thead>
<tr>
<th>TEST value</th>
<th>Hypothyroid</th>
<th>Euthyroid</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 or less</td>
<td>18</td>
<td>1</td>
</tr>
<tr>
<td>&gt; 5</td>
<td>14</td>
<td>92</td>
</tr>
<tr>
<td>Totals</td>
<td>32</td>
<td>93</td>
</tr>
</tbody>
</table>

- Sensitivity is $0.56$ (18/32) and the specificity is $0.99$ (92/93)
• Let’s increase the threshold to 7

<table>
<thead>
<tr>
<th>T4 value</th>
<th>Hypothyroid</th>
<th>Euthyroid</th>
</tr>
</thead>
<tbody>
<tr>
<td>7 or less</td>
<td>25</td>
<td>18</td>
</tr>
<tr>
<td>&gt; 7</td>
<td>7</td>
<td>75</td>
</tr>
<tr>
<td>Totals</td>
<td>32</td>
<td>93</td>
</tr>
</tbody>
</table>

Sensitivity is 0.78 and the specificity is 0.81.

• Let’s increase the threshold to 9

<table>
<thead>
<tr>
<th>T4 value</th>
<th>Hypothyroid</th>
<th>Euthyroid</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 9</td>
<td>29</td>
<td>14</td>
</tr>
<tr>
<td>9 or more</td>
<td>3</td>
<td>39</td>
</tr>
<tr>
<td>Totals</td>
<td>32</td>
<td>93</td>
</tr>
</tbody>
</table>

Sensitivity is 0.91 and the specificity is 0.42.

- Can improve sensitivity by moving threshold to a higher T4 value
- Can improve the specificity by moving threshold to a lower T4 value
- There is a tradeoff between sensitivity and specificity.

**ROC curve**

- Receiver Operating Characteristic curve (or ROC curve.)
- Plot of the true positive rate against the false positive rate for the different possible cutpoints of a diagnostic test.

**Training vs. Testing**

- Need to separate training data from testing data
- Why?
Cross Validation

• Say we decide to use 80% of the data for training and 20%.
• Which 80% should we use?

Cross Validation

• Varies the way the data is partitioned into training and test data.
• N-Fold Cross Validation
  – Each partition is a fold
  – N is the number of times you do this

Cross Validation

• What do you report?
  – “accuracy” measure(s)
  – Stdev
  • Provides a measure of how sensitive your system is to the training data

Recap

• Now you are:
  – Familiar with some different techniques for understanding volumes of data
  – Familiar with some different information retrieval and classifier metrics
  – Are familiar with some NLP methods