Overview of MUC and Introduction to Text Mining

Jakub Piskorski
Feiyu Xu
DFKI LT-Lab
Overview of Message Understanding Conferences
Message Understanding Conferences
[MUC-7 98]

- U.S. Government sponsored conferences with the intention to coordinate multiple research groups seeking to improve IE and IR technologies (since 1987)

- defined several generic types of information extraction tasks (MUC Competition)

- MUC 1-2 focused on automated analysis of military messages containing textual information

- MUC 3-7 focused on information extraction from newswire articles
  - terrorist events
  - international joint-ventures
  - management succession event
Evaluation of IE systems in MUC

- Participants receive description of the scenario along with the annotated training corpus in order to adapt their systems to the new scenario (1 to 6 months)

- Participants receive new set of documents (test corpus) and use their systems to extract information from these documents and return the results to the conference organizer

- The results are compared to the manually filled set of templates (answer key)
Evaluation of IE systems in MUC

- Precision and recall measures were adopted from the information retrieval research community.

\[
\text{recall} = \frac{N_{\text{correct}}}{N_{\text{key}}} \quad \text{precision} = \frac{N_{\text{correct}}}{N_{\text{correct}} + N_{\text{incorrect}}}
\]

- Sometimes an $F$-measure is used as a combined recall-precision score.

\[
F = \frac{(\beta^2 + 1) \times \text{precision} \times \text{recall}}{\beta^2 \times \text{precision} + \text{recall}}
\]
Generic IE tasks for MUC-7

- (NE) Named Entity Recognition Task requires the identification and classification of named entities
  - organizations
  - locations
  - persons
  - dates, times, percentages and monetary expressions

- (TE) Template Element Task requires the filling of small scale templates for specified classes of entities in the texts
  - Attributes of entities are slot fills (identifying the entities beyond the name level)
  - Example: Persons with slots such as name (plus name variants), title, nationality, description as supplied in the text, and subtype.

  “Capitan Denis Gillespie, the commander of Carrier Air Wing 11”
(TR) Template Relation Task requires filling a two slot template representing a binary relation with pointers to template elements standing in the relation, which were previously identified in the TE task:

- subsidiary relationship between two companies
  (employee_of, product_of, location_of)
(CO) Coreference Resolution requires the identification of expressions in the text that refer to the same object, set or activity

- variant forms of name expressions
- definite noun phrases and their antecedents
- pronouns and their antecedents

“The U.K. satellite television broadcaster said its subscriber base grew 17.5 percent during the past year to 5.35 million”

- bridge between NE task and TE task
(ST) Scenario Template requires filling a template structure with extracted information involving several relations or events of interest

- intended to be the MUC approximation to a real-world information extraction problem
- identification of partners, products, profits and capitalization of joint ventures

**Generic IE tasks for MUC-7**

**SOURCE:** Jakub Piskorski, Feiyu Xu Sommersemester 2001
# Tasks evaluated in MUC 3-7

[Chinchor, 98]

<table>
<thead>
<tr>
<th>EVAL\TASK</th>
<th>NE</th>
<th>CO</th>
<th>RE</th>
<th>TR</th>
<th>ST</th>
</tr>
</thead>
<tbody>
<tr>
<td>MUC-3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>YES</td>
</tr>
<tr>
<td>MUC-4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>YES</td>
</tr>
<tr>
<td>MUC-5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>YES</td>
</tr>
<tr>
<td>MUC-6</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td></td>
<td>YES</td>
</tr>
<tr>
<td>MUC-7</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
</tr>
</tbody>
</table>
Maximum Results Reported in MUC-7

<table>
<thead>
<tr>
<th>MEASURE\TASK</th>
<th>NE</th>
<th>CO</th>
<th>TE</th>
<th>TR</th>
<th>ST</th>
</tr>
</thead>
<tbody>
<tr>
<td>RECALL</td>
<td>92</td>
<td>56</td>
<td>86</td>
<td>67</td>
<td>42</td>
</tr>
<tr>
<td>PRECISION</td>
<td>95</td>
<td>69</td>
<td>87</td>
<td>86</td>
<td>65</td>
</tr>
</tbody>
</table>
Introduction to Text Mining
Goal

- Help develop an understanding of what text mining is
- Improve the overview of the applications of text analysis techniques
- Push the further development of the language technologies
What is Text Mining

KNOWLEDGE DISCOVERY
“Extraction of implicit, previously unknown, and potentially useful information from data” [Frawley, 1991]

DATA MINING (KDD)
Extraction of hidden, predictive information from large databases [Zhang and Currim, 2000]

TEXT MINING
Discovery of knowledge in unstructured textual data
What is Text Mining

- Information Extraction
- Information Retrieval
- Data Mining
- Text Categorization
- Computational Linguistics
- Machine Learning
- Probabilistic Modeling

Source: Jakub Piskorski, Feiyu Xu

Sommersemester 2001
What is Data Mining?

[Witten & Frank 2000]

- Data: information locked up in the database which is potentially useful
- Data Mining (Knowledge Discovery from Databases)
  - the extraction of implicit, previously unknown, and potentially useful information from data
    - finding patterns across large datasets
  - building programs that sift through databases automatically, seeking regularities or patterns.
    - discovering heretofore unknown information
  - Technical basis: statistical and machine learning methods
### Example

[Witten & Frank 2000]

<table>
<thead>
<tr>
<th>Width</th>
<th>Height</th>
<th>Sides</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>4</td>
<td>4</td>
<td>standing</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>4</td>
<td>standing</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>4</td>
<td>lying</td>
</tr>
<tr>
<td>7</td>
<td>8</td>
<td>3</td>
<td>standing</td>
</tr>
<tr>
<td>7</td>
<td>6</td>
<td>3</td>
<td>lying</td>
</tr>
<tr>
<td>2</td>
<td>9</td>
<td>4</td>
<td>standing</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>4</td>
<td>lying</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>3</td>
<td>lying</td>
</tr>
</tbody>
</table>
Example

- A possible rule will be produced for this data is
  
  If width \( \geq 4.0 \) and height \( \leq 6.0 \) then lying
  Otherwise if height \( \geq 4.0 \) then standing

- But how about width=1 and height=2?

- Better classification rule
  
  If width \( > \) height then lying
  If height \( > \) width then standing
## Example

<table>
<thead>
<tr>
<th>Width</th>
<th>Height</th>
<th>Sides</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
<td>3</td>
<td>?</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>4</td>
<td>?</td>
</tr>
</tbody>
</table>
Why Data Mining?
[Hearst ACL 99]

- Because the data is there.

- Because
  - larger disks
  - faster cpus
  - high-powered visualization
  - networked information

are becoming widely available.
What is Text Mining?
[Hearst ACL 99]

- Analyzing text to extract implicit information and discover new knowledge that is useful for particular purpose
  - Main subtasks: text clustering, topic detection, ontology generation
  - Complementary areas: information extraction, text categorization

- Compared to data bases, text is unstructured, ambiguous and therefore more difficult to deal with

- Text mining is possible because you do not have to fully understand text in order to extract useful information from it

- Why text mining?
  - There are more unstructured text documents than databases, for example, web documents
How does Text Mining Differ from Information Retrieval?  [Hearst 97, 99]

- The fact that an information retrieval system can return a document that contains the information a user requested does not imply that a new discovery has been made: the information had to have already been known to the author of the text; otherwise the author could not have written it down.
  - ad hoc query
  - The system has available only pre-existing, indexed documents.
  - Its response is limited to selecting from these documents relevant to the queries and presenting them to the user, namely, finding and ranking documents that match users’ information need
  - It must select, say, 10 or 20 documents from millions.
Text Mining vs. Information Retrieval

Text Mining Research Group at the University of Waikato
...
docs. Publications in the Text Mining area by group members. Appeared: ... Waikato, ...
Description: The Text Mining group at the University of Waikato in New Zealand. With a focus on...

IBM Data Management White Papers: Text Mining
...
Description: Tweeting information into knowledge. An IBM white paper. Better utilize electronic mail from customers, ...

Text mining and web-based information retrieval reference
Text Mining, Web Mining, Information Retrieval and Extraction from the WWW References: On ...
Text Mining Application

- Extract patterns that can be incorporated in text document management tools:

  - Knowledge Base
  - Categorization
  - Routing
  - Filtering
  - Segmentation
  - Retrieval
  - Ranking
  - Summarization
  - Clustering
  - Organization
  - Navigation

Source: Jakub Piskorski, Feiyu Xu
Finding patterns across text is already done in CL

Goal: automated language understanding

- This is difficult and almost impossible
- instead, go for subgoals, e.g.,
  - word sense disambiguation
  - phrase recognition
  - semantic associations

Common current approach:

- statistical analyses over very large text collections
  - Part-of-speech tagging [Armstrong1994]
  - WordNet acquisition [Fellbaum1998]
  - Automatic acquisition of subcategorization data from text corpora [Manning 1993]
Why CL Isn’t Text Mining
[Hearst ACL99]

- A linguist finds it interesting that “cloying” co-occurs significantly with “Jar Jar Binks” ...

- … But this doesn’t really answer a question relevant to the world outside the text itself.
Why CL Isn’t Text Mining
[ Hearst ACL 99 ]

- We need to use the text indirectly to answer questions about the world.

Example

⇒ Direct:
- Cluster newswire text; determine which terms are predominant

⇒ Indirect:
- Analyze newswire text; gather evidence about which countries/alliances are dominating which financial sectors
Is CL Useful for Text Mining? And Why?
[Hearst ACL99]

- Text is tough

- CL can do preprocessing for the text mining to clean up the text data
Text is Tough
[Hearst ACL 99]

- Abstract concepts difficult to represent

- “Countless” combinations of subtle, abstract relationships among concepts

- Many ways to represent similar concepts
  - space ship, flying saucer, UFO, figment of imagination

- Concepts are difficult to visualize

- High dimensionality
  - Tens or hundreds of thousands of features
I saw Pathfinder on Mars with a telescope.

Pathfinder photographed Mars.

The Pathfinder photograph mars our perception of a lifeless planet.

The Pathfinder photograph from Ford has arrived.

The Pathfinder forded the river without marring its paint job.
Früher stellten die Frauen der Inseln am Wochenende Kopftücher mit Blumenmotiven her, die ihre Männer an den folgenden Montagen auf dem Markt im Zentrum der Hauptinsel verkauften. (Uszkoreit)

- Lexical and morphological ambiguity (32)
- Morphosyntactic ambiguity (case ambiguity) (8)
- Attachment ambiguity (252)
  - PP attachment (63)
  - Extraposed relative clauses (4)

=64.512 readings
Text is Easy
[Hearst ACL 99]

- Highly redundant in bulk

- Just about any simple algorithm can get “good” results for coarse tasks
  - Pull out “important” phrases
  - Find “meaningfully” related words
  - Create summary from document
Text Data “Clean Up”
[Hearst ACL 99]

- Tokenization
- Morphological Analysis (Stemming)
  inflectional, derivational, or crude IR methods
- Part-of-Speech Tagging
  I/Pro see/V Pathfinder/PN on/P Mars/PN ... 
- Phrase/Sentence Structure Analysis
  [PRO I][VP [V saw] [PN Pathfinder] [PP on Mars] [PP with a telescope]]
- Predicate/Argument Structure
- Scope of Conjunctions
- …
Example: Ontology Acquisition
[Hearst 92, 99]

- **Application of a lexico-syntactic pattern:**
  
  \[
  \text{NP}_0 \text{ such as NP}_1, \{\text{NP}_2, \ldots \ (\text{and} \mid \text{or}) \ \text{NP}_i\} \\
  i \geq 1, \implies \text{forall NP}_i, \ i \geq 1, \ \text{hyponym}(\text{NP}_i, \ \text{NP}_0)
  \]

- **Extracts out a new hypernym:**
  
  Agar is a substance prepared from a mixture of red algae, such as Gelidium, for laboratory or industrial use.

  \implies \text{hyponym}(\text{Gelidium, red algae})
Text Mining vs. Text Categorization
[Hearst 97, 99]

- **Text Categorization**: sorting the specific content of a document into one (or more) of a set of pre-defined classes

- **Questions:**
  - Is text categorization text mining?
  - Can the text categorization results be used for text mining?
  - How much linguistic preprocessing is needed for the improvement of text categorization? [Neumann&Schmeier 99]
Text Mining vs. Topic Detection
[Hearst 99, Allan et al. 98]

- Topic detection is the discovery of the beginning of a new theme or trend by taking a stream of new stories as input and classify them into a new theme or to already discovered themes.

- Topic Detection = Text Mining + Text Classification?
Text Mining vs. Information Extraction
[Nahm 2001]

- The goal of text mining is to discover knowledge in unstructured texts
- The goal of information extraction is to transform unstructured text into a structured database
  ⇒ Information extraction + Data Mining = IE-based Text Mining Framework
Summary

- **Text Mining:**
  - Extracting heretofore undiscovered and useful information from large text collections

- **Related Fields**
  - Data Mining
  - Information Extraction
  - Computational Linguistics
  - Information Retrieval
  - Text Categorization

- **Subtasks**
  - Topic detection
  - Clustering
  - Ontology acquisition