Short Talks about Final Projects

Adam Meyers
New York University
Outline

• Preliminary Schedule of Talks
• Structure of Talks
• Sample Slides
Preliminary Schedule of Talks

• Organized by topic
  – These are rough approximations of topic area. If you believe your talk is in the wrong category, please let me know

• Final Schedule available on Thursday, Dec 3
  – Some talks may be shifted from Tuesday to Thursday or vice versa
Talks on Tuesday December 8

- Document Classification and Detection
  - Burke, Tanner M
  - Kwon, Sean Hyuk Lynn
  - Lage, Erika
  - Li, Shannon

- Document Summarization
  - Bokhnyak, Sergiy Viktorovych
  - Prentice, Kira C

- Generation
  - Tezapsidis, Katerina M
  - Young, Halley R
  - Burns, Terri Christina
  - Mao, Daniel X
  - Masanto, Corey
  - Yenisey, Tolga
Talks on Thursday, December 10

- **Sentiment**
  - Tianco, Rene
  - Drachman, Isaac W
  - Hua, Ye
  - Zhu, Jielei
  - Li, Andrew S
  - Jin, Kevin

- **Information Extraction**: Saxe, Ryan C
- **Question Answering**: Ryan, John P
- **Sense Disambiguation**: Dumas, Maximillian Francis
- **Coreference**: Feldman, Trey
- **Translation**: Ma, Long
Structure of Sessions

• Each Student will have 3 minutes plus 1 minute for questions. I will allow another 1 minute for transitions between talks
  – $5 \times 26 = 130$ minutes
  – 2 classes $= 150$ minutes

• There is an extra 10 minutes/class should some talks go over the time limit, but I will speed things along if necessary.

• For multi-person projects, students can combine their talks and are allowed $6 + 2$ minutes or $7 + 1$. 
Suggested Structure of a 3 minute talk

• Estimate rate of talk: 1 slide per minute.
  – So a 3 minute talk should be about 3 slides in addition to any slides you don't discuss

• Possible structure:
  – Slide 1: title and author(s) – Don't discuss
  – Slide 2: problem statement
  – Slide 3: your approach: methods, resources used (lexicons, corpora, programs incorporated), features of data used, algorithms implemented, evaluation
  – Slide 4: One line conclusion
  – Slide 5: References – Don't discuss
Possible Structure for a 2 Person Project Talk

• Assume either a 6 + 2 or 7 + 1 structure

• Slides
  – Slide 1: title and author(s) – Don't discuss
  – Slide 2: problem statement
  – Slide 3: your approach: methods, resources used (lexicons, corpora, programs incorporated), features of data used, algorithms implemented
  – Slide 4: Evaluation
  – Slide 5: Role of Participant 1
  – Slide 6: Role of Participant 2
  – Slide 7: One line conclusion
  – Slide 8: References – Don't discuss
Other Factors

• Talks in the same topic areas
  – There may be redundancies between talks. You could discuss this and have some of these redundancies be factored out and presented by one person (1 or 2 minutes could be added to the schedule for this). I can provide email addresses if this helps.
  – You could learn from the other talks about resources that will help you with your own project.
• You don't have time to give a detailed account of your approach
• The best talks will communicate the problem and approach to solving it in a conversational manner
• Imagine you explaining this to a non-technical people with short-attentions spans: people at a party, relatives, etc.
  – But perhaps a little more technical than that
• Your final project may change a lot from the one you present as a talk, e.g., you may end up incorporating comments from other students or answering their concerns
Sample 3 + 1 Talk

• Next 5 slides represent a hypothetical talk based on some of my MT slides
Sentence Alignment Using Gale Shapey Stable Marriage Algorithm

Adam Meyers
New York University
Sentence Alignment Problem

- Bitexts = Source Language file and sentence by sentence translation
- Sentences may be in different orders
- In order to “train” statistical systems, we need to know which source language sentence corresponds to which target language sentence

<table>
<thead>
<tr>
<th>English</th>
<th>Spanish</th>
</tr>
</thead>
<tbody>
<tr>
<td>Run John run</td>
<td>Corre Sara corre</td>
</tr>
<tr>
<td>See John run</td>
<td>Ve Sara corre</td>
</tr>
<tr>
<td>Run Sally run</td>
<td>Corre Juan corre</td>
</tr>
<tr>
<td>See Sally run</td>
<td>Ve Juan corre</td>
</tr>
<tr>
<td>Run Spot run</td>
<td>Corre Mancha corre</td>
</tr>
<tr>
<td>See Spot run</td>
<td>Ve Mancha corre</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
My Approach

• Score matches by overlap in a bilingual lexicon (supplement with automatic acquisition of additional entries) using Dice

\[ Dice = \frac{2 \times |\text{Match}(S,T)|}{|S| + |T|} \]

• Compare scores of neighboring sentences in 10 X 10 array
  – Choose best match for first sentence by Gale Shapely algorithm, then advance 10 sentence window

• Gale Shapely maximizes choices between 2 sets of competing items, where there are optimal choices, e.g., like arranging marriages between potential sets of mates

• Corpus: Spanish/English Microsoft Help Text

• Evaluation Metrics: Precision, Recall, F-measure
Selected References


