Short Talks about Final Projects

Adam Meyers
New York University
Outline

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

- Student Talks presented for about 2 classes
  - April 27 and May 2
  - Preliminary Schedule on next 2 slides
- Organized by approximate topic
  - Different talks may be grouped together same group
- Please make schedule changes by email
  - If project is miscellaneous — email me your topic
    - I would like to label and put similar topics together
  - If you need a different day – email me
- Timing different for multi & single person projects
  - Single person: 3 minutes plus 1 minute for questions
  - Multi person: 4 minutes plus 1 minute for questions
  - Extra time allowed for time overages and transitions
Please Send Me Your Slides

• Please send me your slides
  – pdf format preferred
  – please send at least one day before your talk
  – this saves time and makes the talks go smoother in general

• You can send me updated versions later

• You can use your own laptop if you prefer
Talks on Thursday, April 27
Total Time: 46 minutes

- Parsing (4 min)
  - C Abraham
- Machine Translation (4+5+4+4=17 min)
  - E Galvis
  - G Csapo and J Kim
  - Z Zhao
  - H Yang
- Text Simplification (4 min)
  - A Ramachandran
- Automatic Essay Scoring (5+4=9 min)
  - K Rosenberg and P Sukhum
  - Z Mo
- Dialogue (4+4=8 min)
  - J White
  - M Salinas
- Miscellaneous (4×4=16 min)
  - A Gold
  - R Krishnan
  - S Malkowicz
Talks on Tuesday May 2
Total Time: 63 min

• Deep Learning (4 min)
  – R Cruz
• Sentiment (5+5+(5x4)=30 min)
  – I Perfito and S Undavia
  – J Denker, J Zhang, Y Li, and J Yao
  – R Hidalgo
  – S Wang
  – J Camacho
  – J Peng
  – T Masri
• Summarization (4 min)
  – J Woods
• Document Classification (5+4+4=13 min)
  – E Chan, M Kim, S Mun and Y Kim
  – K Zhang
  – M Hinderer
• POS tagging (4 min)
  – P Sereiko
• Miscellaneous (4 min)
  – A Puri
  – V Aggarwal
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 Multi Person Project Talk

• 4 minutes or 1 Minute Per Slide
• Sample 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
  – Slide 4: Roles of Each Participant
  – Slide 5: One line conclusion
  – Slide 6: 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.
    • Project mergers are permitted
    • Other types of coordination are permitted, e.g., a shared evaluation for different methods
  – 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

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

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

• Gale Shapey maximizes choices between 2 sets of competing items, where there are optimal choices, e.g., like a arranging marriages between potential sets of mates
Evaluation and Conclusion

- Corpus: Spanish/English Microsoft Help Text
- Evaluation Metrics: Precision, Recall, F-measure
- Manually aligned Microsoft Help bitext for development
  - 1350 English and 1341 Spanish Sentences
- Manually aligned bitext used for testing
  - 184 English and 181 Spanish sentences
- We achieved F-measure of 95%
Selected References


