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 26 and May 1
  – 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
Talks on Thursday, April 26
Total Time: 65 minutes

• Document Classification (30 min)
  – M. Vu (3 + 1 min)
  – J. Shang (3 + 1 min)
  – N. Spektor (3 + 1 min)
  – L. Leung (3 + 1 min)
  – M. Ahmed (3 + 1 min)
  – O. Giddens and Z. Kimelheim (4+1 min)
  – B. Peng and J. Wang (4 + 1 min)

• Sentiment Analysis (21 min)
  – C. Dickstein (3 + 1 min)
  – D. Jin and C. Li (4 + 1 min)
  – N. Marshall (3 + 1 min)
  – J. Young (3 + 1 min)
  – J. Lee (3 + 1 min)

• Other (14)
  – C. H. Lee and N. Oberman (4 + 1 min)
  – V. Aggarwal and A. Puri (4+ 1)
  – A. D'sa (3 + 1)
Talks on Tuesday, May 1
Total Time: 66 minutes

- Information Extraction (20 min)
  - E. Paik (3 + 1 min)
  - C. Tong (3 + 1 min)
  - R. Abramson (3 + 1 min)
  - B. Diaz (3 + 1 min)
  - M. Meserve (3 + 1 min)

- Machine Translation (9 min)
  - M. Brekke, N. Gheorghe, C. Lopez (4 + 1 min)
  - J. Dorville (3 + 1 min)

- Question Answering (8 min)
  - M. Dinku (3 + 1 min)
  - I. Kulkarni (3 + 1 min)

- Document Summarization (8 min)
  - Z. Guo (3 + 1 min)
  - M. Ligier (3 + 1 min)

- Annotation (4 min)
  - A. Isenberg (3 + 1 min)

- Segmentation/POS Tagging (4 min)
  - A. Wang (3 + 1 min)

- Language Identification (4 min)
  - P. Merritt (3 + 1 min)

- Miscellaneous (9 min)
  - A. Yang and L. Shi (4+1)
  - M. Yang (3+1)
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
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-attention 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

```
Run John run    Corre Sara corre
See John run    Ve Sara corre
Run Sally run   Corre Juan corre
See Sally run   Ve Juan corre
Run Spot run    Corre Mancha corre
See Spot run    Ve Mancha corre
...              ...
```
My Approach

- Score matches by overlap in a bilingual lexicon (supplement with automatic acquisition of additional entries) using Dice
  \[
  Dice = \frac{2 \times |Match(S, T)|}{|S| + |T|}
  \]
- 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 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


