Short Talks about Final Projects

Adam Meyers
New York University
Outline

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

• Student Talks presented for 2 ½ classes
  – April 27, May 2 and May 4
• Organized by topic
  – These are rough approximations of topic area. Very different talks can be in the same group
• Final Schedule available on Thursday, April 25
  – Some talks may be shifted by 1 day
• 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
Talks on Wednesday, April 27

- Sentiment Detection (12 talks)
  1. Ameen A Ajmeri
  2. Anhad S. Arora
  3. Cameron L. Ballard & Marshall R. Evans
  4. Abhinav Hvaldar & Pauline Ayumi Okuda Ceraulo
  5. Jhishan Khan
  6. Katherine M Herrick & Andrew C Klingelhofer & Deniz E Parlar
  7. Mark M Joseph & Derek Dp Yu & Amanda K Sewanan
  8. Savannah S Lim
  9. Monica Martinez
  10. Alexandra V Serralta & Jiachen Zhao
  11. Dexter Y Tong
  12. Sanjna Verma

- Information Extraction (3 talks)
  13. John Sampson
  14. Sana Sheikh & Jessica M Wang
  15. Zbynek Stara
Talks on Monday, May 2

• Generation and Text Augmentation (6 Talks)
  – Christopher V Elwood
  – Dennis McDaid & Carter Yu
  – Arvind Ramgopal & Michael L Tan
  – Yanqiu Wu
  – Kenneth Y Zhang
  – Danny O Vilela

• Summarization (6 Talks)
  – Meet Manoj Barot & William T Kim
  – Dhruv Gupta & Shantanu Joshi
  – Carlos Guzman & Nicholas Jackson
  – Leslie Johanna Manrique
  – Baron I Nzekwe
  – Clara Rubin

• HMM (2 Talks)
  – Mengmei (Ivy) Chen
  – Milivoje Divovic

• Machine Translation (2 Talks)
  – Gina Elbert
  – Sabreen Noah
Talks on Wednesday, May 4

• Information Retrieval and Document Classification
  – Li, Jason
  – Edgar Montano & Jeremie Roche
  – Max Lebedev
  – Aviv R Goldgeier

• Finite State Machine/Regular Expressions: George Miller

• Annotation:
  – Daniel Balagula
  – Sarah Lerner

• 2 Other Talks
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 Shapley 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
... ...

See John run
Run Sally run
See Sally run
Run Spot run
See Spot run
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 Shapley algorithm, then advance 10 sentence window

• Gale Shapley 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


