Final Project Discussion

Adam Meyers
Montclair State University
Summary

• Project Timeline
• Project Format
• Details/Examples for Different Project Types
  – Linguistic Resource Projects: Annotation, Lexicons, ...
  – Programming Projects
  – Research Paper
  – Other
• Note: I will give mostly English examples, but you can use any language for your data: systems that process other languages, resources for other languages, etc. are all appropriate.
Project Time Line

• Project Proposal
  – Due November 17

• Project Talk
  – Presented on December 1 or 8

• Final Project Report Due
  – December 15
Format

• Use the ACL style information either for Word or latex for ACL 2014
  – Go to the call for papers:
    • http://acl2014.org/CallforPapers.htm
    • Scroll down to “ACL 2014 Style Files” and use the appropriate style files (latex or Word)
  – Please include your name on the paper and do not try to make it anonymous as the ACL instructions indicate (i.e., you are not submitting the paper for blind review)
• The final papers should be approximately 6-10 pages, not including references.
  – The page length is less important than the content
  – You may hand in other materials in addition to the paper
Proposal (due Nov 17)

• A couple pages showing that your paper is plausible
• Problem Statement or Introduction
  - Indication about what motivates the work you are describing: the research question you are investigating, the purpose of the algorithm you are implementing, etc.
• Strategy for Solving the Problem
  - **Annotation Project**: initial specifications, a small amount of annotation (e.g., 1 page), and a plan on how to achieve a modest amount of high-quality annotation.
  - **Lexicon Project**: initial specifications, a small number of lexical entries, a word list or method for deriving word list, method for deriving dictionary (manual, automatic, semi-automatic)
  - **Programming Project**: Description of an existing algorithm and how you plan to implement it and test it.
  - **Research paper**: what you plan to read and how you expect it will address the problem statement
Baseline and Full Systems

• When describing the results it is often useful to compare these results to some baseline, a system used for purposes of comparison.

• The baseline system can represent an obvious method that any system should beat
  – For example, for POS taggers, choose most frequent tag for each word based on training corpus, and choose NN for all unknown words.

• It is possible that such a system could get 75% accuracy (suppose 50% of tokens have only 1 possible tag and another 25% have one tag that makes up at least 50% of the instances).

• Another (higher) baseline system could be also a proof-of-concept system, the very basic implementation of what you are doing before adding all the bells and whistles.

• Similar to a program, annotation guidelines may include distinctions that are more or less difficult tags to annotate consistently. A baseline system may collapse difficult distinctions. For example, suppose you are tagging adverbs and your system distinguishes the 2 instances of the adverb logically listed below. It may make sense to evaluate annotator agreement with and without this distinction.
  – Logically, they should be able to do it. [VIEWPOINT]
  – They should be able to solve the problem logically [MANNER]

• It is a good idea to mention one or more baselines in your proposal. It is a good idea to guarantee that you can produce something simple, even if your more advanced ideas interest you more.
Linguistic Annotation Projects

• Write Specifications & Annotate Documents
  – Find related work and distinguish your approach
  – Find one annotator in addition to yourself, so that it is possible to measure inter-annotator agreement
    • Multiply annotate and adjudicate sample for evaluation purposes
    • Or multiply annotate and adjudicate for all annotation
  – Possible deals:
    • annotate for each other
    • Programmer uses annotation for program project

• Design and Implement Annotation Project using Amazon Turk
  – Figure out a way to use very simple annotation
  – Design a task for Amazon Turk and Run it
  – May cost you a little money ($50 buys a lot of annotation)
  – If anyone is interested in this route, I can provide more info
Sample Annotation Projects

- Apply one a known type of Annotation (NE, POS, Chunking, semantic role labeling) to a new domain of text: web data, technical data, a new language, etc.
- Develop specifications and annotate new classes of NEs, Relations, or Events
- Develop specifications and annotate an interesting phenomenon:
  - quantifier scope
  - sentiment (your version)
  - Idiomatic expressions
Lexicon Projects

• Motivation
  – What would your lexicon help achieve?
  – How would you test this?

• Strategy for Constructing Lexicon
  – Word List:
    • I can provide large lists of English lemmas (and morphological variants)
    • Lists can be derived from corpora, sorted by frequency
  – Automatic Methods:
    • Even heuristics that are 50% accurate can save time, e.g., for finding a list of place names, this pattern could help: `grep -E 'going to [A-Z]'`
  – Lexicographer Interfaces (may be programs online)
    • Bad idea to have lexicographers type in features
  – Write up and Test Specifications: similar to annotation
  – Evaluation
    • Consistency (like annotation)
    • Demonstrate that information in the lexicon can be used for some application
      – Demonstrate that it would case the correct prediction most of the time.
Existing & Future Lexicons

• Existing Lexicons and Databases
  – Comlex Syntax: Syntactic & Semantic
  – Nomlex-Plus, ADJADV: Paraphrase
  – WordNet: Word Sense
  – CIA FactBook: Gazetteer

• New Genres Where Lexicons Could Help: Twitter, etc.
Sample Programming Projects

• Sequence Labeling System
  – Strategies: Rule based, HMM or Other
  – Types: POS tagger, Chunker, NE tagger, Time Expression, ...

• Phrase Structure Generation Program
  – Change the system to make more plausible sentences
  – Generate Poetry

• Implementing Existing NLP Algorithm, even if not covered in class
Survey Papers

• If you select an area, I can recommend some articles/books to get you started

• Research Areas: Machine Translation, Summarization, Question Answering, Sentiment Analysis, Information Extraction, Reference Resolution, Predicate Argument Structure, …

• Your proposal should include an abstract, a description of some preliminary work that you have read, a list of articles that you intend to read and an enumeration of your research goals.
Resources Available

- In addition to publications, there are other resources available to help with your research.
  - Corpora
  - Lexicons
  - Programs

- If you are creating a resource or program, you could possibly use these resources as pre-processors.
  - If you are classifying adjectives, a POS tagger may help you find adjectives
  - A parser would be a good pre-processor to a system doing pronoun coreference
The Wall Street Journal Penn Treebank

• I could make this available in several forms
• Upenn: Trees, POS tags, Noun Groups
• BBN:
  – NE – Inline annotation – you can convert it to one token per line or use it as is
  – Coreference – marks pronoun coreference – there is probably a bit of corpus preparation to make this work
• There are some licensing issues, so I have to distribute these in a careful way.
Downloadable Tasks with Annotated Corpora (for Testing and/or Training)

• Corpora for Drug-Drug Interaction
  – http://labda.inf.uc3m.es/DrugDDI/DrugDDI.html

• WePS – searching for entities on the Web

• Spanish Corpus with POS tags
  – http://www.comp.lancs.ac.uk/linguistics/crater/spanishfiles.html
Some downloadable corpora

- The Open American National Corpus
  - http://www.anc.org/OANC/
  - A variety of different types of data
  - A limited amount of manually annotated data
  - Automatically annotated data from various programs
  - Most annotation is offset annotation

- The Singapore SMS corpus

- Wikipedia XML

- Tweet Corpus (for sentiment)
  - http://help.sentiment140.com/for-students/
Some Lexicons

- COMLEX – licensing issues, but I could make it available
- NOMLEX and related dictionaries
    - “Those Other Dictionaries” describe the dictionaries
    - “directory linked here” brings you to directory of dictionaries and other resources
    - Everything as one archive file (Nombank 1.0 Release)
- WordNet: http://wordnet.princeton.edu/
- FrameNet: https://framenet.icsi.berkeley.edu/fndrupal/
- CMU Pronunciation Dictionary: http://www.speech.cs.cmu.edu/cgi-bin/cmudict
- Subjectivity Lexicon (and sentiment corpus)
  - http://mpqa.cs.pitt.edu/
- CIA factbook:
Tools and Packages

• You can use output of NLP tools, as input to your system, provided you make your contribution clear.

• Examples:
  – NLTK
  – Machine Learning Packages (as per talks)
  – Other tools you download
Please Ask

• If you need help finding resources or citations relevant to your project, please send me email.