

# Hands-On NLP for an Interdisciplinary Audience

**Elizabeth D. Liddy & Nancy J. McCracken**  
**Center for Natural Language Processing**  
**School of Information Studies**  
**Syracuse University**

# Overview

- **Interdisciplinary Course**

- Need a **single** course that meets needs of divergent audience
  - Multi-disciplinary
    - Information Science (5 degree programs), Computer Science, Linguistics
  - Multiple levels – PhD, masters, a few undergrads
    - + Commercial & government folks
- 600 level course in the School of Information Studies
- 3 hr class, once a week, for 15 weeks

- **Goals**

- Provide students with a broad, solid basis on which to build
  - Theory and basic computational linguistic principles
  - Hands-on implementations
  - Range of possible applications
- Enable students to decide whether to pursue more detailed understanding

# Focus of Paper / Presentation

- Person-centered aspects of designing / conducting course
  - Human & social needs
  - Essential for making course with mixed expertise work well
  - Contented & challenged students make for a good learning and teaching experience
- Experience shows you need be attentive to structuring course to ensure that students:
  - Are comfortable both with what they know and relying on others for what they don't know – AND –
  - Are willing to risk extending themselves

# Topics

- Organized around the levels of language processing & computational techniques for each
  - Morphology    Finite state automata
  - Lexicology    Part-of-speech tagging
  - Syntax        Parsing with context free grammars
  - Semantics     Word sense disambiguation
  - Discourse     Sublanguage analysis
  - Pragmatics    Gricean Maxims
- Readings
  - Jurafsky & Martin's textbook
  - Seminal & recent papers

# Course Organization

- Designed around multiple types of team projects
  - Membership changes for each project
    - Teaches students with diverse backgrounds to value divergent experience
    - Advances learning further than if working alone or with the same team throughout the course
    - Forms a class that thinks of itself as a community
- Frequent, short presentations of group work
  - Enables students to own their new understandings
  - Replaces all written papers
  - Learn from work of other groups
  - Students learn to explain and defend their decisions

# Hands on Features

1. Small, in-class group simulations of computational processes.
2. Team projects implementing various levels of NLP using open-source software on large corpora.
3. Team posters and public presentations on the state-of-the-art in commercial applications such as:
  - summarization
  - machine translation
  - speech recognition
  - text mining
  - question answering
  - language generation

# In-class Group Simulations – 1

- Part of each intro lecture on lower levels of language processing
- Paper & pencil exercises that simulate the process they have just learned
  - Part-of-speech tagging
  - Parsing simple sentences with small grammar
- Needed because not everyone understands how an algorithm will work

# In-class Group Simulations – 2

- New ad hoc groups of 4 students each time
  - Formed by professor
  - Ensure that each student works with every other student early on in the course
- Supports learning
  - Breaks down social barriers and makes it easier for students to share their work
  - Students learn to value other discipline's contribution
    - Computer science  $\leftrightarrow$  linguistics
    - Technical  $\leftrightarrow$  analytical
  - Simulations help students better understand the process they are about to automate

# Text Processing Projects

- Goals for students
  - Gain hands-on experience in utilizing NLP software
  - Analysis of a large, real-world data set
- Project assignment in two parts:
  - Computing POS tags of text
  - Identifying chunked phrases of text
- Use of the NL processed text to characterize the data

# Project Teams

- Teams
  - Mix of computer science, linguistics and information science experience
  - Membership determined by professors based on short survey of students' skills & preferences
- Organized tasks within teams by interest & ability
  - Choosing a data analysis application
  - Processing data subsets
  - Designing NL processing to accomplish the analysis
  - Programming the NL processing
  - Conducting the data analysis
  - Preparing in-class reports

# Corpus

- Enron email corpus
  - <http://www-2.cs.cmu.edu/~enron/>
  - Contains 250,000 unique email messages
    - Among 500,000 messages, 2.75 gigabytes
  - Email folders of 150 people
  - Each team chose a subset of email to work on
- Email formatting software
  - Adapted from Andrés Corrada-Emmanuel's of UMass, Amherst

# NL Toolkit

- Used lower levels of NLP from the NL Toolkit (Loper & Bird 2002)
  - Core data types for NLP
  - Statistical processing
  - Standard NLP algorithms for
    - Tokenization, POS tagging, chunking
    - Tokenizers and chunkers with user-defined regular expressions
    - (In future, would add parsing or other more advanced levels)
- Ran 2 lab sessions for students to get familiar with NLTK and Python
  - Used NLTK tutorials and documentation
  - Prepared additional examples

# Student Projects – Part I

- POS tagging
  - Used regular expression tokenizers to tailor tokenization to email text
  - Trained and used POS taggers from either Brown or Penn Treebank corpora
- Example project: Disambiguate mentions of people by first name
  - Used POS tagged text to characterize text by noun frequencies
  - Compared frequency distribution of emails of all “Toms” to decide which “Tom” is mentioned in one email

# First Name Disambiguation

## *E-mail body:*

... Every thing is good here. **Tom** and I played golf yesterday at Fox Hollow, beautiful day, unfortunately can't say the same for our golf game. Enron is sure in the press. Seems like your hotdog McKinnsey ex-president and CFO are much more adept at accounting sleights of hand than actually running the company.

## *Count frequencies of nouns in the E-mail body:*

<i>1. matthew</i>	<i>2</i>	<i>5. october</i>	<i>2</i>
<i>2. contract</i>	<i>2</i>	<i>6. acceptance</i>	<i>1</i>
<i>3. golf</i>	<i>2</i>	<i>7. administration</i>	<i>1</i>
<i>4. hal</i>	<i>2</i>	<i>8. article</i>	<i>1</i>

# Compare Frequencies

## Tom Brady :

1. *bobby*
2. *knight*
3. *net*
4. *pat*
5. *mike*

6. *b*
7. *bob*
8. *david*
10. *joseph*

## Tom Lester:

1. *fantasy*
2. *game*
3. *smith*
5. *player*
6. *sunday's*
7. *johnson*
8. *numbers*
10. *football*

**Golf game X 1 match Tom Lester "game" X 86**

## Tom Martin :

1. *week*
2. *practice*
3. *game*
4. *league*
5. *reports*
6. *san*
7. *season*
8. *sunday*
9. *james*
10. *squad*

# Student Projects – Part II

- Students used regular expressions to implement a variety of chunk types
  - Noun phrase, verb phrase, maximal noun phrase
- Example student project
  - Analysis of social status of Bill Williams with respect to people he corresponds with
  - Patterns of communication verbs and other phrases

# Examples of Phrases Identified

## BOSS INDEX

I like that  
I [^ ]\* looking (more )?for  
This (should |will )?works?  
Please (send|make|check|fix)  
(screw-up|wrong|hurt)  
Well done  
(must|should) be  
Could you  
We have [^ ]\* positions? open  
We (just )?need to make sure  
ASAP  
Thank you for the update  
I need you to  
make sure you were aware  
please come see me  
bring your resume  
ha(d|ve|s) worked for me  
works? fo me  
(be|make) sure

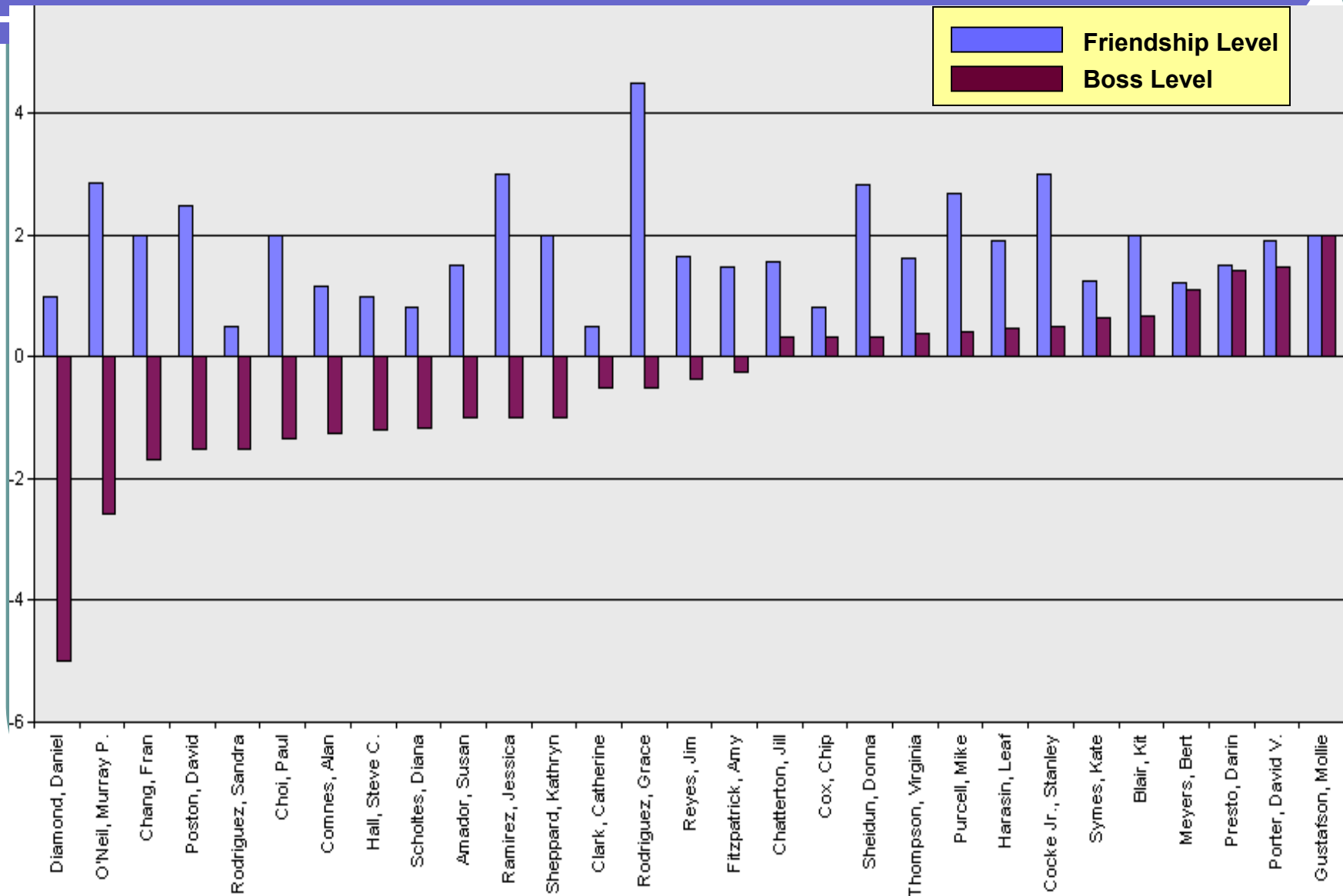
## DEPENDENT INDEX

Let me know if you have any  
other questions  
Let me know if you need  
anything else  
^Here'(s|re)  
^Here (is|are)  
my (duties|responsibilities)  
Thank you for (the|this)  
opportunity  
(I will|I'll) be available  
Sorry  
Do you want  
request your approval  
you hiring me  
work for you

## EQUAL INDEX

(I would|I'd) like to help you  
Check this out  
Thanks  
what ever  
Come on  
Ok  
I got you a ticket  
Sweet  
awesome  
(Let me know if)? (there)? (is)?  
anything I can help with  
I am sorry to hear  
If you are interested in  
FYI  
Howdy  
could you help me with  
hell  
golf

# Frequencies of Phrase Types



# Team Posters & Public Presentations - 1

- Goal: Understand how NLP is utilized in state-of-the-art commercial applications
- Motivation
  - Information Management Masters need to advise employers on language-based applications
  - All students appreciate gaining new in-depth understanding of leading edge technologies
  - PhD students identify needed research to improve or extend existing applications

# Team Posters & Public Presentations - 2

- Students supplied with potential applications:
  - Machine Translation
  - Text Mining
  - Summarization
  - Spell Correction
  - Cross-Language Retrieval
  - Question Answering
  - Speech Recognition
  - NL Generation
  - Information Retrieval
  - Dialogue Agents
- Given 2 weeks to familiarize themselves with applications
- Sign up on a first-come basis for 4-person teams
- Three 10 - 20 minute report backs over the semester
- Students must actively critique other teams on both content and presentation (counts toward grade)

# Team Posters & Public Presentations - 2

- Students supplied with potential applications:
  - Machine Translation
  - Text Mining
  - Summarization
  - Spell Correction
  - Cross-Language Retrieval
  - Question Answering
  - Speech Recognition
  - NL Generation
  - Information Retrieval
  - Dialogue Agents
- Given 2 weeks to familiarize themselves with applications
- Sign up on a first-come basis for 4-person teams
- Three 10 - 20 minute report backs over the semester
- Students must actively critique other teams on both content and presentation (counts toward grade)

# Team Posters & Public Presentations - 3

- Content guidelines for each presentation
  1. Non-technical overview of generic application + examples of publicly available systems
  2. Technical details, concentrating on computational linguistic aspects used in application
  3. Combines best of 1 & 2 in a poster plus demo of a product in an open school-wide presentation

# Why is Poster Reception important?

- Gives students more than a single chance to produce their best explanation
- Feedback from faculty & other students serves to broaden & deepen their knowledge
- Students take great pride in sharing their new knowledge & expertise
- Wider exposure of NLP applications builds an audience for future semesters
- Instills in faculty & students a sense of the reach and importance of NLP

# Evaluation - 1

- Students by both professors
  - In-Class group exercises 20%
  - NLToolkit Assignments 35%
  - Application Presentations & Poster 35%
  - Contributions to class discussion 10%
    - Both quality and quantity
    - Including commenting on other teams' presentations

# Evaluation - 2

- Peer evaluation by other team members & self
  1. Text processing team
  2. NLP applications team
- Evaluate
  - Role / tasks of the student
  - Overall performance rating (1 to 4)
  - Rationale for this score
  - How the student could improve
- Same grade is assigned to all team members unless peers' input dictates otherwise

# Evaluation - 3

- Professor by student
  - Mid-term & end-of-semester
  - Quantitative scores
    - Ability to communicate
    - Familiarity with content
    - Availability
  - Open-ended questions
    - What worked well in this course?
    - What didn't work well?
    - What the professor could do to improve?

# Evaluation Comments

- **Want:**
  - Easier text book
  - Just 1 semester long project – OR –
  - Same team for both projects
  - Optional primer class session on Linguistics
  - Incorporate WordNet so they can do more semantic level work
- **Liked:**
  - Real world applications
  - Field trip reports from conferences & project meetings
- **More of:**
  - Classroom exercises
  - Labs (for some)

# Indicators of Success

1. Whether students take another NLP course
  - Information Retrieval, Data Mining
2. Decision by undergrads and masters to go on to an advanced degree with NLP focus
3. PhD students continue research in NLP
  - Focus of / part of dissertation
  - Research Assistantships in CNLP
4. Terminal degree students seek & obtain jobs that utilize NLP
5. Hearing each student population pick up & use the vocabulary of the other population

# Best Indicator of Success

- PhD students requested a new seminar course
  - Y.DOT (Your Data Our Tools)
- Those doing social science research want to explore how they can use NLP in their work
- Students will bring their own text data sets
  - Interview transcripts
  - Chatrooms, blogs
  - Open-ended questions on surveys

# Y.DOT

- Each will present their data & hypotheses, run data, experiment, and analyze
- CNLP folks will work with them to specialize and adapt our NLP tools
- Interleave NLP tools with capabilities of current commercial *content analytic* software on their data
  - Adapt generic Information Extraction to tags based on researcher's model or code book
- Keen interest from both PhD students and faculty who will bring their data and sit in
  - Public policy, communications, management, info science
- Basis of recent NSF proposal on assisting social scientists to do “big science” with NLP + CA tools