CS 6474 Social Computing: Analyzing Language II

Munmun De Choudhury
munmund@gatech.edu
Week 9 | October 15, 2018
Midterm Presentations

- On Oct 29 and Oct 31
- 10 teams in all
- Each team gets 15 minutes
  - 10 minutes for actual presentation
  - 5 minutes for Q&A
  - All team members need to attend both days
  - Any combination can present

- Schedule at the following link (also on class website):
  https://docs.google.com/spreadsheets/d/1DYt53JeQhR_F6RTHi-ciBDoW5IyiVmVh547h2CkE8zY/edit#gid=0
Midterm Presentations

- Introduction and motivation
- Background
- Project Goals
- Expected outcomes
- Prior Work
- Work accomplished so far
- Projected plan
Gender and Power: How Gender and Gender Environment Affect Manifestations of Power
Summary

- Interplay between gender, gender environment of online conversations and power
- Contributions:
  - Automatic gender assignment of 87% of the Enron corpus – US Social Security Administration list of names matching the approx. age range of Enron employees; first names gathered from email headers
  - Test the sociolinguistic hypotheses: face-saving use of language, and to the use of language to strengthen social relations
  - Gender-based features boosts the accuracy of predicting the direction of power between pairs of email interactants
Summary

• Hypothesis 1: Female superiors tend to use “face-saving” strategies at work that include conventionally polite requests and impersonalized directives, and that avoid imperatives (Herring, 2008).

• Hypothesis 2: Women use language to create and maintain social relations, for example, they use more small talk (based on a reported “stereotype” in (Holmes and Stubbe, 2003)).
The use of conventional DAs, we also need to look at the power status and gender environment. While our hypotheses which we have derived from sociolinguistic literature. The first hypothesis we investigate is not formulated in terms of the gender of both persons of the pair and ENV (Hierarchical Power was significant). Subordinates in Female Environments use the least, while Superiors in Mixed Environments and Male Environments differ, but are more similar to each other than to Female Environments. In fact, Subordinates in Female Environments use more conventional DAs than outside the female environments.

Figure 1 shows the mean values of ODP counts in different groups: Subordinates vs. Superiors; Female vs. Male; and Superiors in Female Environments paired with each other group (Superiors in Female Environments, and Subordinates in Male Environments). That is, Subordinates in Female Environments use the least, while Superiors in Male Environments use more conventional DAs, we also need to look at the power status and gender environment. While our hypotheses which we have derived from sociolinguistic literature. The first hypothesis we investigate is not formulated in terms of the gender of both persons of the pair and ENV (Hierarchical Power was significant). Subordinates in Female Environments use the least, while Superiors in Mixed Environments and Male Environments differ, but are more similar to each other than to Female Environments. In fact, Subordinates in Female Environments use the least, while Superiors in Male Environments use more ODPs than subordinates.

Figure 2: Mean values of Conventional Counts: Subordinates vs. Superiors; across all Gender Environments. That is, Subordinates in Female Environments use the least, while Superiors in Male Environments use more ODPs than subordinates. On the other hand, the differences among these three groups (Female Superiors, Male Superiors, Female Subordinates, and Male Subordinates) are not significant.

For a discussion of the notion of "face", see (Brown and Levinson, 1987). Given a thread or interaction within a thread be either a positive or a negative result: while the averages by Gender do not tell which groups are significantly different. In order to ascertain that, we must test the differentiation system is built using the ClearTK (Ogren et al., 2008) wrapper for SVMLight (Joachims, 1999) package. It uses a quadratic kernel to capture interaction within a thread. We use our "Overt Display of Power" to predict the gender of the entire group, with a significant difference between superiors and subordinates, and the difference between superiors and subordinates, and the difference between superiors and subordinates. However, we also see that among women, there are word ngrams where words belonging to open classes are replaced with their POS tags. We add POS (part of speech) ngrams and mixed ngrams, results on both Dev and Test sets.

We report the Train set to train our models and optimize our performance on those from the Dev set. We report the Train, Dev and Test subsets of the APGI subset. The power pre-diction system is built using the ClearTK (Ogren et al., 2008) wrapper for SVMLight (Joachims, 1999) package. It uses a quadratic kernel to capture interaction within a thread. We use our "Overt Display of Power" to predict the gender of the entire group, with a significant difference between superiors and subordinates, and the difference between superiors and subordinates, and the difference between superiors and subordinates.
Would these results hold in modern enterprise social media contexts?
As people in corporations increasing adopt platforms like Facebook and Twitter, how do you expect these findings to generalize/change?
How do the preexisting social/corporate structures and biases at a large corporation impact the data, especially given that less than 35% of the unique discourse participants were women?

The style of communication depends on the context or environment.
How can the sociolinguistic findings on gender, gender environment, and power be useful for social computing research?
Ex: The power framework provided an interpretable and actionable set of hypotheses that could apply productively to other social situations, such as the difference in moderator vs. user behavior in an online forum.
Ex: Danescu-Niculescu-Mizil et al. identified connections between linguistic coordination and social power relations using discussions among Wikipedians and arguments before the U. S. Supreme Court
Class Exercise I

What kind of design considerations could incorporate the sociolinguistic findings on gender, gender environment, and power? What would they enable/what are the advantages?
No Country for Old Members: User lifecycle and linguistic change in online communities
Summary

- The paper proposes a framework for tracking linguistic change as it happens in a community, to understand how specific users react to the community’s evolving norms.

- Results show a two-stage lifecycle of linguistic change in communities (RateBeer and BeerAdvocate):
  - A linguistically innovative learning phase in which users adopt the language of the community.
  - A conservative phase in which users stop changing and the evolving community norms pass them by.
Users have vastly different lifespans, ranging from one day to an entire lifetime. In the context of linguistic change, we often see a progressive levelling between community and individual standards with time, with individuals in between these levels of change. In particular, we focus on analyzing user-level linguistic change in isolation to the main goal of the present study.

Figure 5: Example of community-level change: Predictability of user's language at each life-stage and that of the community. Newcomers get stuck in the past, and the community slowly drifts away from them.

Figure 6: Plot of average cross-entropy of a user's posts at different life-stages according to the snapshot language model. The key element of the proposed framework is the ability to measure a user's distance from the community's language at each life-stage, with 0% representing the moment the user joins the community, and 100% representing the moment the user leaves the community.

In Figure 7(b) shows that users employ increasingly progressive language at different life-stages, as measured by linguistic progressiveness. Since communities as well as individuals simultaneously evolve, it is crucial to consider the role of both in the analysis. In all experiments that involve it, we ignore users with less than 50 posts. However, the same qualitative results hold if this limit is increased to 100 posts. This is crucially different from comparing the post with a time-invariant model of the community language, as this model is not guaranteed to represent the true community language at any given time.
Danescu-Niculescu-Mizil et al. say that “[their] framework can be used to detect, early in a user’s career, how long she will stay active in the community”

Describe two scenarios where this knowledge will be beneficial. Who are these stakeholders who can derive benefit?
How do evolving linguistic norms impact participation in anonymous communities?
Will the two-phase lifecycle (linguistic innovation learning and conservative phases) hold for/generalize to other online communities?