Understanding Differences in Written Correspondence Between Superiors and Subordinates

Nicholas Vogler
University of California, Berkeley School of Information
nvogler@berkeley.edu
July 2018

Abstract

E-Mail is the primary communication method used in corporations across the globe. Employees at all levels of an organization use this platform to connect with other employees at the organization across the ranks. This system provides a structure that allows for an analysis of the difference in the language used to communicate to senior and junior employees. This project models these interactions with deep learning. Subject lines, footers, and the bodies of e-mails are grouped to represent a message. A multi-layer long short-term memory recurrent neural network model will take a message as input and return a prediction of the relationship between the sender and the receiver. This research is carried out using the Enron e-mail dataset of traffic between 148 employees and their job titles. {TBD comment on results}

Introduction (motivation for work)

Is the language used to communication down an organization hierarchy different than language used to communicate up? The goal of this research is to both answer that question and determine how that language differs. Enron's collapse in 2002 led to the public dissemination of the largest dataset of email traffic within a single corporation available. Language used throughout this traffic is candid, natural and precisely represents the shape and structure of written communication in a corporate environment. It is within that structure that we can explore the many intricacies of professional relationships.

Understanding how members of different roles within an organization communicate could be useful in a variety of domains outside of the corporate environment. Marketing firms would benefit applying similar models on social networks to construct hierarchies where none formally exist. Information security personnel could use this information to help determine when an account has been compromised. Follow-on research could be conducted to explore how 'communicating like a leader' affects an individual's career in the long-term.

In this research, a LSTM RNN is built and tested against a Multinomial Naïve Bayes baseline and previous research. {Further expand on results briefly mentioned in the Abstract}

Background (literature review, or related work)

Given the quality and size of this dataset, extensive research has been carried out on it since its release. A subset of this research studies the power dynamics and social relationships as discussed in this paper. Many early studies used an additional dataset mapping employee names to job titles to create truth data for the organizational hierarchy.

Research completed by Gilbert (2012) defined these mappings. Bramsen (2011) and Prabhakaran (2014) used similar methods of determining truth in their work, creating SVMs to provide their most accurate

results. Bramsen compiled all correspondence between two individuals and used n-grams within the collection as whole to determine relationships. His approach resulting in a 78.9% accuracy. Prabhakaran evaluated e-mails on an individual basis using n-grams as well, but also included additional non-lexical features. He compiled those results to form a conclusion on relationships, returning an accuracy of 73.03%.

Apoorv Agarwal introduced a gold standard by that attempts to provide truth data for power relationships, gender, and employee roles across the entire organization. However, this truth data is not publicly available. If this data cannot be acquired prior to the conclusion of this research, truth data will instead be extracted from the job titles of 108 employees.

Research by Kong et al. included all of the approaches above in addition to a variety of neural network models to answer. Their study determined that a convoluted neural network CNN built on grouped e-mails as done in Bramsen's research with additional non-lexical features as included in Prabhakaran's research resulted in 81.8% accuracy. Their CNN model outperformed all other NN implementations in every combination of these approaches, and returned more accurate results than both Bramsen and Prabhakaran in their respective studies.

Methods (include a description of any proposed work even if not completed yet)

A data set of job titles paired with employee names was joined to the corpus of emails to identify the position of the sender and receiver. This was done by pairing the UID of a sender (the primary key in the job titles data set) to the email address most frequently used to send from that UID, where the UID is listed in the x-origin section of the email. Messages sent from distribution groups were retained, but were excluded from this count.

Of the 148 employees included in the Enron e-mail dataset, only 108 of them were identified in the dataset of job titles. The 40 unlisted employees were removed from consideration. For employees with multiple titles, the higher title is used. The hierarchy of the remaining identified employees is shown on the left in figure 1, with the hierarchy used by Gilbert shown on the right. Both hierarchies are used in separate tests, the latter for comparison.

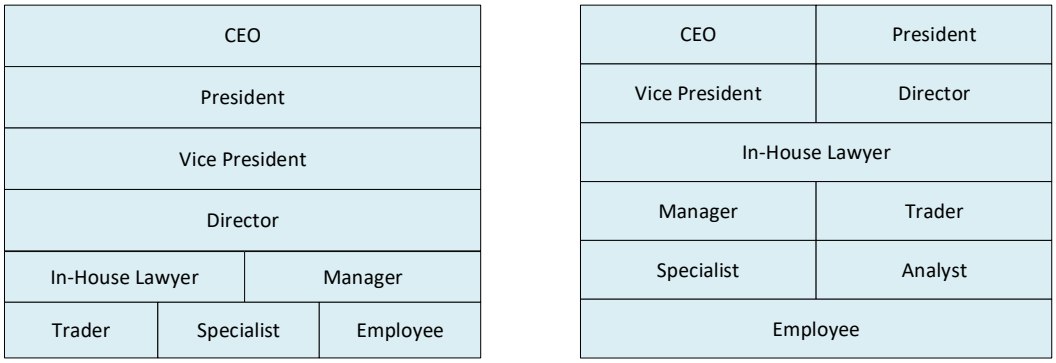

Figure 1. Organizational Hierarchies including all the job titles of Enron Employees in the dataset

Messages sent to individuals above the sender are labeled upward. Messages sent to individuals below the sender are labelled not-upward. Messages that do not fall into these categories are neutral.

Each record in the dataset contains the file name and the entire contents of the e-mail. Employees are identified in the 'To' and 'From' sections of the e-mail and verified against the job titles mappings. If the e-mail meets the criteria for either of the labels above, the subject line is joined to the body of the e-mail and added to the set of correspondence for that sender/receiver pair. CC'd employees are not considered. The e-mail is evaluated for each employee in the 'To' section independently. Content of forwarded messages was replaced with the single word 'Forwarded'.

With the messages joined and transformed into a single entity for each relationship, stop words and non-alphanumeric characters are removed. The remaining content is stemmed and tokenized. Models take the resulting data structure as input.

After data preparation was complete, there were a total of X relationships. Y labelled upwards, Z labelled not-upward, and A neutral.

Results and discussion (for your baseline model, though feel free to include material for anything else you've done)

719 distinct sender:receiver pairs were identified such that both the sender and the receiver's job title was known. Of these 719, 256 are labelled upward and 238 are labelled downward. The remaining 225 are neutral.

To establish a baseline, a Multinomial Naïve Bayes classifier was built with the SkLearn platform. Without any optimizations, the results were only slightly better 50/50 coming in at ~52% accuracy.

(Results and Conclusions) Next Steps section for work you plan to do before submitting the final version (you'll remove this section and replace it with your conclusions, final results and analysis in your final report)

With the pipeline established and the baseline model ran, the next step is to address the inefficiencies in the data preparation. The process of identifying senders and receivers needs to be verified for all messages to ensure as much of the data that can be used is. Aliases need to be merged into a single entity. Contents of forwarded e-mails need to be added as new messages to account for more of the available data.

Follow on steps will be to implement and optimize an LSTM RNN and compare results to the baseline and similar studies.

From here, additional hierarchy structure will be tested and further refinement of the data preparation process will be completed.

References

Gilbert, Eric. "Phrases that signal workplace hierarchy." *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. ACM, 2012.

Bramsen, Philip, et al. "Extracting social power relationships from natural language." *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 2011.

Prabhakaran, Vinodkumar, and Owen Rambow. "Predicting power relations between participants in written dialog from a single thread." *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Vol. 2. 2014.

Agarwal, Apoorv, et al. "A comprehensive gold standard for the enron organizational hierarchy." *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*. Association for Computational Linguistics, 2012.

Kong, Angela. Lam, Michelle. Xu, Catherina, “Power to the People: Using Deep Learning to Predict Power Relations.” 2015. Print.