Identifying Dominance Relationships in Written Correspondence

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

Abstract

Email 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 convoluted neural network model will take a message as input and return a prediction of the relationship between the sender and the receiver. The model This research is carried out using the Enron email dataset of traffic filtered down to conversations between employees with identified job titles.

Introduction

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, LSTM-CNN, and CNN models are built and tested against a Multinomial Naïve Bayes baseline and previous research. None of the neural network models were found to outperform the baseline.

Background

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. His research focused on creating a hierarchy of the employees and their positions. The associations between job titles and employees were extracted from his data set and used as a source of truth for employee ranking in this research.

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

An entity map pairing employee emails to job titles was joined to the corpus of emails to identify the hierarchical rank of the sender and receiver. Content of each email was added to the conversation from a sender and a receiver. An email sent to multiple employees results in additions to the conversation between the sender and each of the receivers individually. Messages sent from distribution groups or employees of the same rank were retained, but were excluded from this count.

After filtering the data set to include only employees with known job titles communicating to only other employees with known job titles, the initial 36027 relationships identified were reduced to 10906. With the remaining conversations, 192538 emails were sent. An email sent to two employees was considered two distinct emails.

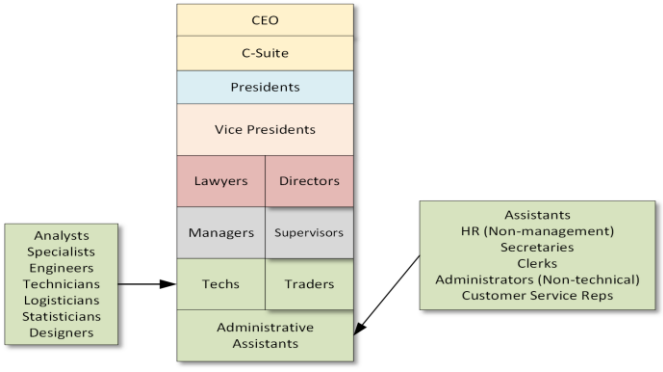

Figure 1. Organizational Hierarchies including all the job titles of Enron Employees in the dataset

Messages sent to employees above the sender are labeled upward. Messages sent below the sender are labelled down. Messages that do not fall into these categories are neutral and were not used in any models.

When all conversations were formed, stop words and numeric characters are removed. Remaining content is further filtered to remove non-informational content and tokenized. Models then take the resulting data structure as input with a 70%/30% split between training and development.

After data preparation was complete, 6154 employee email addresses were associated with 368 distinct job titles. Each employee may be mapped to multiple e-mail addresses.

Results

A Multinomial Naïve Bayes classifier built with unigrams was used as a baseline. Without any optimizations this model correctly identified 55% of the conversation's relationships. By adding n-grams the accuracy was improved by 8%, and was more accurate than a simple feed forward neural network.

Models	Accuracy	
	Raw Email Content	After Additional Cleansing
Unigram Naive Bayes	55%	63%
N-Gram Naive Bayes	63%	65%
Feed Forward Neural Network	61%	59%

Figure 2. Baseline model results on the initial dataset and the same models ran after further data cleansing was completed.

Feature analysis identified severe flaws in the data preparation process, and led to additional cleansing. This cleansing improved the accuracy on the Multinomial Naïve Bayes models but reduced the accuracy of the baseline neural network as shown in Figure 2. LSTM, CNN, and CNN-LSTM models were built using the cleansed revision of the data. These models did not perform better than the baseline as shown in Figure 3.

Model	Accuracy
LSTM	60%
CNN	57%
CNN-LSTM	52%

Figure 3. Final model results with accuracy as the metric.

Failure to improve upon the baseline results was expected to be caused by the limited size of usable data. An excerpt from a misclassified conversation, *adobe acrobat read print documents adobe acrobat reader downloaded free http adobe com,* reveals a flaw in the data preparation process allowing advertisements to contribute to the classification of conversations. Further discoveries identified employees used various methods of forwarding emails which were not accounted for, erroneously keeping them inconsideration. Additional tuning of the data preparation process is expected to greatly improve results.

Optimization of the baseline Multinomial Naïve Bayes N-Gram had slight improvement increasing the accuracy by 1%.

Conclusion

Key phrases were identified that indicated a conversation was more likely to be sent from a higher ranking employee to a lower ranking employee. The phrases most important to this conclusion were *let know, would like, pleas let, pleas let know, and staff meet* as generated by the Multinomial Naïve Bayes n-gram model. However, the confidence in such a classification is limited due to limited accuracy of the models. Improvement the data preparation process, and therefore increasing the quality and quantity of the data being used to create the model, is expected to greatly improve results.

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.

Namata, Galileo Mark S., Lise Getoor, and Christopher P. Diehl. "Inferring organizational titles in online communication." *Statistical Network Analysis: Models, Issues, and New Directions*. Springer, Berlin, Heidelberg, 2007. 179-181.

Heer, Jeffrey. "Exploring Enron: Visualizing ANLP results." *Applied Natural Language Processing*. 2004.