Detecting Power Relations from Written Dialog

Vinodkumar Prabhakaran
Department of Computer Science
Columbia University
New York, NY 10027, USA
vinod@cs.columbia.edu

Abstract

In my thesis I propose a data-oriented study on how social power relations between participants manifest in the language and structure of online written dialogs. I propose that there are different types of power relations and they are different in the ways they are expressed and revealed in dialog and across different languages, genres and domains. So far, I have defined four types of power and annotated them in corporate email threads in English and found support that they in fact manifest differently in the threads. Using dialog and language features, I have built a system to predict participants possessing these types of power within email threads. I intend to extend this system to other languages, genres and domains and to improve it's performance using deeper linguistic analysis.

1 Introduction

Social relations like power and influence are difficult concepts to define, but are easily recognizable when expressed. Most classical definitions of power in the sociology literature (e.g. (Bierstedt, 1950; Dahl, 1957)) include "an element indicating that power is the capability of one social actor to overcome resistance in achieving a desired objective or result" (Pfeffer, 1981). Influence closely resembles power, although some consider it as one of the means by which power is used (Handy, 1985). The five bases of power — Coercive, Reward, Legitimate (Positional), Referent, and Expert — proposed by French and Raven (1959) and its extensions are widely used in sociology to study power. I find these definitions and typologies helpful as general background, but not specific enough for a data-oriented study on how they are expressed in online written dialogs.

One of the primary ways power is manifested is the manner in which people participate in dialog. Power relations sometimes constrain how one behaves when engaging in dialog; in some other cases, they enable one to constrain someone else's behavior. And in some cases, the dialog behavior becomes a tool to express and even pursue power. By dialog behavior, I mean the choices one makes while engaging in dialog. It includes choices with respect to the message content, like lexical choices, degree of politeness or instances of overt display of power such as orders or commands. It also includes choices participants make in terms of dialog structure, like the choice of when to participate with how much and what sort of contribution, how many questions to ask and which of those questions to answer and the time between those questions and their answers.

The primary goal of my thesis is to show that different social power relations manifest themselves in written dialog in different, but predictable ways, and to investigate how these manifestations differ across languages, genres and domains. To achieve this goal, I aim to introduce a new typology of power that is relevant in online written interactions and can be validated using data-oriented approaches. Then, I aim to study how these different types of power differ in their manifestations in dialog. Specifically, I aim to capture and compare these manifestations in two dimensions of the dialog: content and structure. In addition to using existing components like dialog act taggers and linkers to capture the dialog structure

Proceedings of the 2012 Student Research Workshop, pages 7–12,

Jeju, Republic of Korea, 8-14 July 2012.

and lexical analyzers to capture content features, I plan to identify and extract more structural and linguistic indicators of power relations. Using these features, I will build a system that can automatically extract power relations between participants of written dialogs across different languages (English vs. Arabic), genres (discussion forums vs. emails) and domains (political vs. scientific). Currently, I have partially achieved this goal within the context of English corporate email threads, which represent a specific language-genre-domain combination. The four types of power I have defined are: situational power, hierarchical power, control of communication and influence. My future research directions include 1) broadening this work onto other languages, genres and domains and 2) using deeper analysis to identify more indicators of power and capture power relations at finer granularity

2 Literature survey

It has long been established that there is a correlation between dialog behavior of a discourse participant and how influential she is perceived to be by the other discourse participants (Bales et al., 1951; Scherer, 1979; Ng et al., 1995). Specifically, factors such as frequency of contribution, proportion of turns, and number of successful interruptions have been identified as being important indicators of influence. Locher (2004) recognizes "restriction of an interactant's action-environment" (Wartenberg, 1990) as a key element by which exercise of power in interactions can be identified. I use a linguistic indicator *Overt Display of Power* which captures action-restriction at an utterance level. Wartenberg (1990) also makes the important distinction between two notions of power: power-over and power-to. Power-over refers to hierarchical relationships between interactants, while power-to refers to the ability an interactant possesses (may be temporarily) and can use within the interaction. My notions of hierarchical power and situational power roughly correspond to Wartenberg's notions of power-over and power-to, respectively. Both can be considered special cases of French and Raven (1959)'s notion of legitimate power. I consider influence as a type of power which captures notions of expert power and referent power described by French and Raven. Finally, my notion of control of communication is based on the concept of conversational control introduced by Ng and Bradac (1993). It is a form of power the participant has over the interaction; other forms of power are modeled between participants.

In computational literature, several studies have used Social Network Analysis (Diesner and Carley, 2005; Shetty and Adibi, 2005; Creamer et al., 2009) to deduce social relations from online communication. These studies use only meta-data about messages: who sent a message to whom and when. For example, Creamer et al. (2009) find that the response time is an indicator of hierarchical relations; however, they calculate the response time based only on the meta-data, and do not have access to information such as thread structure or message content, which would actually verify that the second email is in fact a response to the first.

Using NLP to analyze the content of messages to deduce power relations from written dialog is a relatively new area which has been studied only recently (Strzalkowski et al., 2010; Bramsen et al., 2011; Peterson et al., 2011). Using knowledge of the organizational structure, Bramsen et al. (2011) create two sets of messages: messages sent from a superior to a subordinate, and *vice versa*. Their task is to determine the direction of power (since all their data, by construction of the corpus, has a power relationship). They approach the task as a text classification problem and build a classifier to determine whether the set of all emails (regardless of thread) between two participants is an instance of up-speak or down-speak. In contrast, I plan to use a complete communication thread as a data unit and capture instances where power is actually manifested. I also plan to study power in a broader sense, looking beyond power attributed by hierarchy to other forms of power. Strzalkowski et al. (2010) are also interested in power in written dialog. However, their work concentrates on lower-level constructs called *Language Uses*, which might indicate higher level social constructs such as leadership and power. This said, one of their language uses is agenda control, which is very close to our notion of conversational control. They model it using notions of topic switching, using mainly complex lexical features. Peterson et al. (2011) focuses on formality in Enron email messages and relates it to social distance and power.

3 Work done so far: Power in Corporate Emails

So far, I have worked on my primary goal – studying manifestations of social power relations – within the context of English corporate email threads. For this purpose, I used a subset of email threads from a version of the Enron email corpus (Yeh and Harnly, 2006) in which messages are organized as threaded conversations. In the remainder of this section, I first introduce the power typology and annotations and then present the linguistic and structural features I used. Then, I present the findings from a statistical significance study conducted between these features and different types of power. Finally, I present a system built using these features to predict participants with power within an email thread.

Power Typology and Annotations: After careful analysis of a part of the email corpus, I defined a power typology to capture different types of power relevant in corporate emails. I propose four types of power: situational power, hierarchical power, control of communication and influence.[1] Person_1 is said to have **situational power (SP)** over person_2 if person_1 has power or authority to direct and/or approve person_2's actions in the current situation or while a particular task is being performed, as can be deduced from the communication in the current thread. Person_1 with situational power may or may not be above person_2 in the organizational hierarchy (or there may be no organizational hierarchy at all). Person_1 is said to have **hierarchical power (HP)** over person_2 if person_1 appears to be above person_2 in the organizational hierarchy, as can be deduced from the communication in the given thread (annotators did not have access to independent information about the organizational hierarchy). Possible clues to HP include (by way of example): 1) characteristic of a part of a message as being an approval, or being a direct order; 2) a person's behavior such as asking for approval; 3) a person's authority to make the final decision. A person is said to have **control of the communication (CNTRL)** if she actively attempts to achieve the intended goals of the communication. These are people who ask questions, request others to take action, etc. and not people who simply respond to questions or perform actions when directed to do so. A thread could have multiple such participants. A person is said to have **influence (INFL)** if she 1) has credibility in the group, 2) persists in attempting to convince others, even if some disagreement occurs, 3) introduces topics/ideas that others pick up on or support, and 4) is a group participant but not necessarily active in the discussion(s) where others support/credit her. In addition, the influencer's ideas or language may be adopted by others and others may explicitly recognize influencer's authority.[2] Prabhakaran et al. (2012a) presents more details on annotations of these power relations in the email corpus.

Manifestations in Content and Stucture: I used six sets of features to explore manifestations of power: dialog act percentages (*DAP*), dialog link counts (*DLC*), positional (*PST*), verbosity (*VRB*), lexical (*LEX*) and overt display of power (*ODP*). The first four sets of features relate to the whole dialog and its structure while the last two relate to the form and content of individual messages. The email corpus I used has been previously annotated with dialog acts and links by other researchers (Hu et al., 2009). I used these annotations to capture *DAP* and *DLC* features. *DAP* captures percentages of each of the dialog act labels (Request Action, Request Information, Inform, Conventional, and Commit) aggregated over all messages sent by the participant within the thread. The dialog links include forward links which denote utterances with requests for information or actions, backward links which denote their responses and secondary forward links which denote utterances without explicit requests that were interpreted as requests and were linked back from later utterances. *DLC* captures various features derived from these links with respect to each participant such as counts of each type of link, counts of forward links that are connected back and counts and percentages of those which were not connected back. *PST* includes features to indicate relative positions of first and last messages by a participant. *VRB* includes features to denote how much and how often a participant took part in the conversation. *PST* and

[1] This typology is an extension of an initial typology formulated through collaborative effort with another student.

[2] I adopt this definition from the IARPA Socio-Cultural Content in Language (SCIL) program, where many researchers participating in the SCIL program contributed to the scope and refinement of the definition of a person with influence.

VRB are readily derivable from the email threads. I used simple word ngram features to capture *LEX*.

Overt display of power (*ODP*) is a linguistic indicator of power I introduced. An utterer can choose linguistic forms in her utterance to signal that she is imposing constraints on the addressee's choice of how to respond, which go beyond those defined by the standard set of dialog acts. For example, if the boss's email is "Please come to my office right now", and the addressee declines, he is clearly not adhering to the constraints the boss has signaled, though he is adhering to the general constraints of cooperative dialog by responding to the request for action. I am interested in these additional constraints imposed on utterances through choices in linguistic form. I define an utterance to have *ODP* if it is interpreted as creating additional constraints on the response beyond those imposed by the general dialog act. An *ODP* can be an order, command, question or even a declarative sentence. The presence of an *ODP* does not presuppose that the utterer actually possess social power: the utterer could be attempting to gain power. In (Prabhakaran et al., 2012b), I present a system to identify utterances with *ODP* using lexical features like word and part of speech ngrams along with dialog acts of the utterance.

Statistical significance study: For each type of power, I considered two populations of people who participated in the dialog – $\mathcal{P}_p$, those judged to have that type of power and $\mathcal{P}_n$, those not judged to have that power. Then, for each feature, I performed a two-sample, two-tailed t-test comparing means of feature values of $\mathcal{P}_p$ and $\mathcal{P}_n$. I found many features which are statistically significant, which suggests that power types are reflected in the email threads. I also found that the significance of features differ considerably from one type of power to another, which suggests that these power types are reflected differently in the threads, and that they are thus indeed different types of power. For hierarchical power, the feature TokenRatio has a mean of 0.38 for $\mathcal{P}_p$ and 0.54 for $\mathcal{P}_n$ with a p-value of 0.07. This suggests that bosses tend to talk less within a thread. People with situational power or control request actions significantly more often than others and send significantly more and longer messages than others. People with influence never request actions and send much longer messages than others. They also tend to have more secondary forward links (with a p-value of 0.07) which suggests that people often respond to what people with influence say even if the influencer's contribution is not a request.

Predicting Persons with Power: I formally defined the problem as: given a communication thread $\mathcal{T}$ and an active participant $\mathcal{X}$, predict whether $\mathcal{X}$ has power of type $\mathcal{P} \in \{$SP, HP, INFL, CNTRL$\}$ over some person $\mathcal{Y}$ in the thread. I built a binary SVM classifier for each power type $\mathcal{P}$ predicting whether or not $\mathcal{X}$ has power $\mathcal{P}$ based on features with respect to $\mathcal{X}$ in the context of the given thread $\mathcal{T}$. I obtained good results for SP and CNTRL, but HP and INFL were hard to predict since they occurred rarely in my corpus. The combination of *DLC* and *OSP* performed best for SP (F = 64.4) and *PST* performed best for CNTRL (F = 90.0). For HP, the combination of *DLC* and *LEX* performed best (F = 34.8). For INFL, the best performer was *DLC* (F = 22.6). All results except the ones for INFL were statistically significant improvement over an always-true baseline. I found dialog features to be significant in predicting power, though content features also contribute to detecting some types of power.

4 Proposed Work

So far, I have defined four types of power and have studied how they are expressed and revealed in Enron email threads. My future research directions include deepening this study by i) capturing more linguistic indicators of social power in dialog, ii) building automatic taggers for all linguistic indicators, iii) using deeper semantic analysis on the content and iv) extending it to capture power relations at finer granularity. I also intend to broaden this work into different languages, genres and domains, adapting work done in email threads when viable.

More power indicators : I will work on capturing more linguistic indicators of power from dialog. I currently have annotations at the utterance level that capture attempts to exercise power and attempts to influence. I will use these annotations to build systems that can automatically detect them. In addition, I plan to capture linguistic expressions that suggest lack of power such as asking for approvals, permissions etc. or acting overly polite. For this, I will have to add new annotations to the data. I also plan to perform deeper analysis on the content to capture subjectivity — whether someone states more facts than opinions, commitment — whether someone commits to what she says, and the presence of other modalities such as permissions, requirements, desires etc. I plan to use existing work in subjectivity analysis (Wilson, 2008) and commitment analysis (Prabhakaran et al., 2010) for this purpose. For modality analysis, I plan to use previous unpublished work that I participated in.

Fully automated system: I plan to use automatic taggers to extract dialog act and link features and other linguistic indicators of power (like *ODP*), to build a fully automated social power extraction system. Hu et al. (2009) presented a dialog act tagger and link predictor which could be used to extract *DAP* and *DLC*. However, I found their dialog act tagger performs poorly on minority classes such as requests for actions, which are more critical to predict power. Their link predictor obtained an F measure of 35% which makes it unfit to be used in its current form. For *ODP*, I will use the SVM classifier I built, which obtained a best cross validation F measure of 65.8. I plan to improve the performance of the dialog act tagger, the link predictor and the *ODP* tagger using new features and techniques. I plan to use a threshold adjustment algorithm proposed by Lin et al. (2007) to handle the class imbalance problem in dialog act tagger and link predictor (*ODP* tagger already uses this). I will also build automatic taggers for all other linguistic indicators of power discussed above.

Deeper Semantic Analysis I will explore new features derived from deeper semantic analysis to improve performance of the dialog act tagger, the link predictor and the taggers for other indicators of power like *ODP*. In particular, I plan to use semantic information from VerbNet to provide useful abstraction of verbs into verb classes. This will reduce data sparseness, thereby improving the performance of the taggers. In an initial experiment, I found that using VerbNet class name instead of verb lemma improved the performance of *ODP* tagger by a small margin. I did this only for those verbs that belong to a single VerbNet class (hence needing no disambiguation). I will explore ways to disambiguate verbs with multiple VerbNet class assignments and employ this feature in other taggers as well.

Finer granularity of relations: I will enhance the system to predict power relations between pairs of participants. Aggregating features at the participant level is prone to noise. For example, let $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{Z}$ be active participants such that $\mathcal{X}$ has power over $\mathcal{Y}$, who has power over $\mathcal{Z}$. When we aggregate features with respect to $\mathcal{Y}$, we are introducing noise from the part of communication between $\mathcal{X}$ and $\mathcal{Y}$. Extending my work to the person pair level would prevent this noise and provide us with a finer granularity of power relations. Formally, I want to predict if person $\mathcal{X}$ has power $\mathcal{P}$ over person $\mathcal{Y}$, given a communication thread $\mathcal{T}$. My power annotations already capture the recipient (person_2) of power relations which I will use for this purpose.

Language, genre and domain adaptation: I will extend my work in the English email threads to other languages, genres and domains. Specifically, I plan to work on existing data containing Wikipedia discussion threads and political forums in both English and Arabic. Thus, my thesis would include the analysis of power under 5 different language-genre-domain settings. This step will need extensive annotation efforts. I expect that my proposed power typology might need to be refined to capture types of relations in the new genres. Also, I may have to define new linguistic indicators relevant to the new genres or refine the ones I identified for email threads to adapt to the new genres. This would also require me to adapt various subsystems/taggers to capture features such as dialog acts, links, *ODP* etc. to new genres or build new systems.

5 Conclusion

In my thesis, I propose to study how different power relations are manifested in the structure and language of online written dialogs and build a system to automatically extract power relations from them. I have already conducted this study in English email threads and I plan to extend this to other languages, genres and domains.

6 Acknowledgments

This work is supported, in part, by the Johns Hopkins Human Language Technology Center of Excellence. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the sponsor. I thank my advisors Dr. Owen Rambow and Dr. Mona Diab for their valuable guidance and support. I thank Daniel Bauer for useful discussions and feedback on this proposal.

References

Robert F. Bales, Fred L. Strodtbeck, Theodore M. Mills, and Mary E. Roseborough. 1951. Channels of communication in small groups. *American Sociological Review*, pages 16(4), 461–468.

Robert Bierstedt. 1950. An Analysis of Social Power. *American Sociological Review*.

Philip Bramsen, Martha Escobar-Molano, Ami Patel, and Rafael Alonso. 2011. Extracting social power relationships from natural language. In *ACL*, pages 773–782. The Association for Computer Linguistics.

Germán Creamer, Ryan Rowe, Shlomo Hershkop, and Salvatore J. Stolfo. 2009. Advances in web mining and web usage analysis. chapter Segmentation and Automated Social Hierarchy Detection through Email Network Analysis, pages 40–58. Springer-Verlag, Berlin, Heidelberg.

Robert A. Dahl. 1957. The concept of power. *Syst. Res.*, 2(3):201–215.

Jana Diesner and Kathleen M. Carley. 2005. Exploration of communication networks from the enron email corpus. In *In Proc. of Workshop on Link Analysis, Counterterrorism and Security, SIAM International Conference on Data Mining 2005*, pages 21–23.

John R. French and Bertram Raven. 1959. The Bases of Social Power. In Dorwin Cartwright, editor, *Studies in Social Power*, pages 150–167+. University of Michigan Press.

Charles B. Handy. 1985. *Understanding Organisations*. Institute of Purchasing & Supply.

Jun Hu, Rebecca Passonneau, and Owen Rambow. 2009. Contrasting the interaction structure of an email and a telephone corpus: A machine learning approach to annotation of dialogue function units. In *Proceedings of the SIGDIAL 2009 Conference*, London, UK, September. Association for Computational Linguistics.

Hsuan-Tien Lin, Chih-Jen Lin, and Ruby C. Weng. 2007. A note on platt's probabilistic outputs for support vector machines. *Mach. Learn.*, 68:267–276, October.

Miriam A. Locher. 2004. *Power and politeness in action: disagreements in oral communication*. Language, power, and social process. M. de Gruyter.

Sik Hung. Ng and James J. Bradac. 1993. *Power in language : verbal communication and social influence / Sik Hung Ng, James J. Bradac*. Sage Publications, Newbury Park :.

Sik Hung Ng, Mark Brooke, , and Michael Dunne. 1995. Interruption and influence in discussion groups. *Journal of Language and Social Psychology*, pages 14(4),369–381.

Kelly Peterson, Matt Hohensee, and Fei Xia. 2011. Email formality in the workplace: A case study on the enron corpus. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 86–95, Portland, Oregon, June. Association for Computational Linguistics.

Jeffrey Pfeffer. 1981. *Power in organizations*. Pitman, Marshfield, MA.

Vinodkumar Prabhakaran, Owen Rambow, and Mona Diab. 2010. Automatic committed belief tagging. In *Coling 2010: Posters*, pages 1014–1022, Beijing, China, August. Coling 2010 Organizing Committee.

Vinodkumar Prabhakaran, Owen Rambow, and Mona Diab. 2012a. Annotations for power relations on email threads. In *Proceedings of the Eighth conference on International Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Vinodkumar Prabhakaran, Owen Rambow, and Mona Diab. 2012b. Predicting overt display of power in written dialogs. In *Human Language Technologies: The 2012 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Montreal, Canada, June. Association for Computational Linguistics.

K. R. Scherer. 1979. Voice and speech correlates of perceived social influence in simulated juries. In *H. Giles and R. St Clair (Eds), Language and social psychology*, pages 88–120. Oxford: Blackwell.

Jitesh Shetty and Jafar Adibi. 2005. Discovering important nodes through graph entropy the case of enron email database. In *Proceedings of the 3rd international workshop on Link discovery*, LinkKDD '05, pages 74–81, New York, NY, USA. ACM.

Tomek Strzalkowski, George Aaron Broadwell, Jennifer Stromer-Galley, Samira Shaikh, Sarah Taylor, and Nick Webb. 2010. Modeling socio-cultural phenomena in discourse. In *Proceedings of the 23rd International Conference on COLING 2010*, Beijing, China, August. Coling 2010 Organizing Committee.

Thomas E. Wartenberg. 1990. *The forms of power: from domination to transformation*. Temple University Press.

Theresa Wilson. 2008. Annotating subjective content in meetings. In *Proceedings of the Language Resources and Evaluation Conference*. LREC-2008, Springer. AMIDA-85.

Jen-yuan Yeh and Aaron Harnly. 2006. Email thread reassembly using similarity matching. In *In Proc. of CEAS*.