Predicting Overt Display of Power in Written Dialogs

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

Owen Rambow
CCLS
Columbia University
New York, NY 10027, USA
rambow@ccls.columbia.edu

Mona Diab
CCLS
Columbia University
New York, NY 10027, USA
mdiab@ccls.columbia.edu

Abstract

We analyze overt displays of power (ODPs) in written dialogs. We present an email corpus with utterances annotated for ODP and present a supervised learning system to predict it. We obtain a best cross validation F-measure of 65.8 using gold dialog act features and 55.6 without using them.

1 Introduction

Analyzing written dialogs (such as email exchanges) to extract social power relations has generated great interest recently. This paper introduces a new task within the general field of finding power relations in written dialogs. In written dialog, an utterance can represent an overt display of power (ODP) on the part of the utterer if it constrains the addressee's actions beyond the constraints that the underlying dialog act on its own imposes. For example, a request for action is the first part of an adjacency pair and thus requires a response from the addressee, but declining the request is a valid response. However, the utterer may formulate her request for action in a way that attempts to remove the option of declining it ("Come to my office now!"). In so doing, she restricts her addressee's options for responding more severely than a simple request for action would. Our new task is to classify utterances in written dialog as to whether they are ODPs or not. Such a classification can be interesting in and of itself, and it can also be used to study social relations among dialog participants.

After reviewing related work (Section 2), we define "overt display of power" (Section 3) and then present manual annotations for ODP in a small subset of Enron email corpus. In Section 5, we present a supervised learning system using word and part-of-speech features along with features indicating dialog acts.

2 Related Work

Many studies in sociolinguistics have shown that power relations are manifested in language use (e.g., (O'Barr, 1982)). 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. Through ODP we capture this action-restriction at an utterance level. In the computational field, several studies have used Social Network Analysis (e.g., (Diesner and Carley, 2005)) for extracting social relations from online communication. Only recently have researchers started using NLP to analyze the content of messages to deduce social relations (e.g., (Diehl et al., 2007)). Bramsen et al. (2011) use knowledge of the actual organizational structure to 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). Their reported results cannot be directly compared with ours since their results are on classifying aggregations of messages as being to a superior or to a subordinate, whereas our results are on predicting whether a single utterance has an ODP or not.

2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 518–522,
Montréal, Canada, June 3-8, 2012.

3 Overt Display of Power (ODP)

Dialog is successful when all discourse participants show cooperative dialog behavior. Certain types of dialog acts, notably requests for actions and requests for information (questions), "set constraints on what should be done in a next turn" (Sacks et al., 1974). Suppose a boss sends an email to her subordinate: "It would be great if you could come to my office right now". He responds by politely declining ("Would love to, but unfortunately I need to pick up my kids"). He has met the expectation to respond in one of the constrained ways that the request for action allows (other acceptable responses include a commitment to performing the action, or actually performing the action, while unacceptable responses include silence, or changing the topic). However, dialog acts only provide an initial description of these constraints. Other sources of constraints include the social relations between the utterer and the addressee, and the linguistic form of the utterance. Assume our email example had come, say, from the CEO of the company. In this case, the addressee's response would not meet the constraints set by the utterance, even though it is still analyzed as the same dialog act (a request for action). Detecting such power relations and determining their effect on dialog is a hard problem, and it is the ultimate goal of our research. Therefore, we do not use knowledge of power relations as features in performing a finer-grained analysis of dialog acts. Instead, we turn to the linguistic form of an utterance. Specifically, the utterer can choose linguistic forms in her utterance to signal that she is imposing further constraints on the addressee's choice of how to respond, constraints 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. We are interested in these additional constraints imposed on utterances through choices in linguistic form. We define an utterance to have **Overt Display of Power** (ODP) if it is interpreted as creating additional constraints on the response beyond those imposed by the general dialog act. Note that use of polite language does not, on its own, determine the presence or absence of an ODP. Furthermore, the presence of an ODP does not presuppose that the utterer actually possess social power: the utterer could be attempting to gain power.

ID	Sample utterance
s1	If there is any movement of these people between groups can you please keep me in the loop.
s2	I need the answer ASAP, as ....
s3	Please give me your views ASAP.
s4*	Enjoy the rest of your week!
s5	Would you work on that?
s6*	... would you agree that the same law firm advise on that issue as well?
s7*	can you BELIEVE this bloody election?
s8	ok call me on my cell later.

Table 1: Sample utterances from the corpus; * next to ID denotes an utterance without an ODP

Table 1 presents some sample utterances chosen from our corpus (the * indicates those without ODP). An utterance with ODP can be an explicit order or command (s3, s8) or an implicit one (s2, s5). It can be a simple sentence (s3) or a complex one (s1). It can be an imperative (s3), an interrogative (s5) or even a declarative (s2) sentence. But not all imperatives (s4) or interrogatives (s6, s7) are ODPs. s5, s6 and s7 are all syntactically questions. However, s5's discourse function within an email is to request/order to work on "that" which makes it an instance of ODP, while s6 is merely an inquiry and s7 is a rhetorical question. This makes the problem of finding ODP in utterances a non-trivial one.

4 Data and Annotations

For our study, we use a small corpus of Enron email threads which has been previously annotated with dialog acts (Hu et al., 2009). The corpus contains 122 email threads with 360 messages, 1734 utterances and 20,740 word tokens. We trained an annotator using the definition for ODP given in Section 3. She was given full email threads whose messages were already segmented into utterances. She identified 86 utterances (about 5%) to have an ODP.[1] In

[1]These annotations were done as part of a larger annotation effort (Prabhakaran et al., 2012). The annotated corpus can be obtained at http://www.cs.columbia.edu/∼vinod/powerann/.

order to validate the annotations, we trained another annotator using the same definitions and examples and had him annotate 46 randomly selected threads from the corpus, which contained a total of 595 utterances (34.3% of whole corpus). We obtained a reasonable inter annotator agreement, κ value, of 0.669, which validates the annotations while confirming that the task is not a trivial one.

5 Automatic ODP Tagging

In this section, we present a supervised learning method to tag unseen utterances that contain an ODP using a binary SVM classifier. We use the tokenizer, POS tagger, lemmatizer and SVMLight (Joachims, 1999) wrapper that come with ClearTK (Ogren et al., 2008). We use a linear kernel with $C = 1$ for all experiments and present (P)recision, (R)ecall and (F)-measure obtained on 5-fold cross validation on the data. Our folds do not cross thread boundaries.

5.1 Handling Class Imbalance

In its basic formulation, SVMs learn a decision function f from a set of positive and negative training instances such that an unlabeled instance x is labeled as positive if $f(x) > 0$. Since SVMs optimize on training set accuracy to learn f, it performs better on balanced training sets. However, our dataset is highly imbalanced ($\sim$ 5% positive instances). We explore two ways of handling this class imbalance problem: an instance weighting method, *InstWeight*, where training errors on negative instances are outweighed by errors on positive instances, and *SigThresh*, a threshold adjusting method to find a better threshold for $f(x)$. For *InstWeight*, we used the *j* option in SVMlight to set the outweighing factor to be the ratio of negative to positive instances in the training set for each cross validation fold. *InstWeight* is roughly equivalent to oversampling by repeating positive instances. For *SigThresh*, we used a threshold based on a posterior probabilistic score, $p = Pr(y = 1|x)$, calculated using the ClearTK implementation of Lin et al. (2007)'s algorithm. It uses Platt (1999)'s approximation of p to a sigmoid function $P_{A,B}(f) = (1 + exp(Af + B))^{-1}$, where A and B are estimated from the training set. Then, we predict x as positive if $p > 0.5$ which in effect shifts the threshold for $f(x)$ to a value based on its distribution on positive and negative training instances.

Experiment	InstWeight			SigThresh		
	P	R	F	P	R	F
ALL-TRUE	5.0	100.0	9.5	5.0	100.0	9.5
RANDOM	5.7	58.1	10.4	5.7	58.1	10.4
WORD-UNG	43.1	29.1	34.7	63.0	39.5	48.6
PN,MN,FV,DA	**66.7**	**48.8**	**56.4**	72.3	54.7	62.3
PN,MN,DA	64.5	46.5	54.1	**75.8**	**58.1**	**65.8**
LN,PN,MN,FV	64.4	44.2	52.4	**65.2**	**50.0**	**56.6**

Table 2: Results

Class Imbalance Handling: InstWeight: Instance weighting and SigThresh: Sigmoid thresholding
Features: WORD-UNG: Word unigrams, LN: Lemma ngrams, PN: POS ngrams, MN: Mixed ngrams, FV: First verb, DA: Dialog acts

5.2 Features

We present experiments using counts of three types of ngrams: lemma ngrams (*LN*), POS ngrams (*PN*) and mixed ngrams (*MN*).[2] Mixed ngram is a restricted formulation of lemma ngram where open-class lemmas (nouns, verbs, adjectives and adverbs) are replaced by POS tags. E.g., for the utterance *s2*, *LN* would capture patterns {i, need, i need, ...}, while *PN* would capture {PRP, VBP, PRP VBP, ...} and *MN* would capture {i VBP the NN, ...}. We also used a feature (*FV*) to denote the first verb lemma in the utterance. Since ODPs, like dialog acts, constrain how the addressee should react, we also include Dialog Acts as features (*DA*). We use the manual gold dialog act annotations present in our corpus, which use a very small dialog act tag set. An utterance has one of 5 dialog acts: RequestAction, RequestInformation, Inform, Commit and Conventional (see (Hu et al., 2009) for details). For example, for utterance *s2*, *FV* would be 'need' and *DA* would be 'Inform'.[3]

5.3 Results and Analysis

We present two simple baselines — ALL-TRUE, where an utterance is always predicted to have an ODP, and RANDOM, where an utterance is predicted at random, with 50% chance to have an ODP. We also present a strong baseline WORD-UNG,

[2] *LN* performed consistently better than word ngrams.

[3] We also explored other features including the number of tokens, the previous or following dialect act, none of which improved the results and. We omit a detailed discussion for reasons of space.

which is trained using surface-form word unigrams as features. ALL-TRUE and RANDOM obtained F scores of 9.5 and 10.4 respectively, while WORD-UNG obtained an F score of 34.7 under *InstWeight*, and improved it to 48.6 under *SigThresh*.

For *LN*, *PN* and *MN*, we first found the best value for n to be 1, 2 and 4, respectively. We then did an exhaustive search in all combinations of *LN*, *PN*, *MN*, *FV* and *DA* under both *InstWeight* and *SigThresh*. Results obtained for best feature subset under both configurations are presented in Table 2 in rows 3 and 4. *SigThresh* outweighed *InstWeight* in all our experiments. (Combining these two techniques for dealing with class imbalance performed worse than using either one.) In both settings, we surpassed the WORD-UNG baseline by a high margin. We found *MN* and *DA* to be most useful: removing either from the feature set dropped the F significantly in both settings. We obtained a best F score of 65.8 using *PN*, *MN* and *DA* under the *SigThresh*.

Following (Guyon et al., 2002), we inspected feature weights of the model created for the last fold of our best performing feature configuration as a post-hoc analysis. The binary feature *DA:RequestAction* got the highest positive weight of 2.5. The top ten positive weighted features included patterns like *you_VB*, *_VB*, *MD_PRP*, *VB_VB* and *_MD*, where * denotes the utterance boundary. *DA:Inform* got the most negative weight of -1.4, followed by *DA:Conventional* with -1.0. The top ten negative weighted features included patterns like *MD_VB*, *VB_you*, *what*, *VB_VB_me_VB* and *WP*. In both cases, DA features got almost 2.5 times higher weight than the highest weighted ngram pattern, which reaffirms their importance in this task. Also, mixed ngrams helped to capture long patterns like "please let me know" by *VB_VB_me_VB* without increasing dimensionality as much as word ngrams; they also distinguish *VB_you* with a negative weight of -0.51 from *VB_me* with a positive weight of 0.32, which pure POS ngrams couldn't have captured.

5.4 Not Using Gold Dialog Acts

We also evaluate the performance of our ODP tagger without using gold *DA* tags. We instead use the DA tagger of Hu et al. (2009), which we re-trained using the training sets for each of our cross validation folds, applying it to the test set of that fold. We then did cross validation for the ODP tagger using gold dialog acts for training and automatically tagged dialog acts for testing. However, for our best performing feature set so far, this reduced the F score from 65.8 to 52.7. Our best result for ODP tagging without using gold *DAs* is shown in row 5 in Table 2, 56.9 F score under *SigThresh*. The features used are all of our features other than the *DA* tags. On further analysis, we find that even though the dialog act tagger has a high accuracy (85.8% in our cross validation), it obtained a very low recall of 28.6% and precision of 47.6% for the *RequestAction* dialog act. Since *RequestAction* is the most important feature (weighted 1.7 times more than the next feature), the DA-tagger's poor performance on *RequestAction* hurt ODP tagging badly. The performance reduction in this setting is probably partly due to using gold *DAs* in training and automatically tagged *DAs* in testing; however, we feel that improving the detection of minority classes in dialog act tagging (*RequestAction* constitutes only 2.5% in the corpus) is a necessary first step towards successfully using automatically tagged *DAs* in ODP tagging.

6 Conclusion

We have introduced a new binary classification task on utterances in dialogs, namely predicting Overt Display of Power. An ODP adds constraints on the possible responses by the addressee. We have introduced a corpus annotated for ODP and we have shown that using supervised machine learning with gold dialog acts we can achieve an F-measure of 66% despite the fact that ODPs are very rare in the corpus. We intend to develop a better dialog act tagger which we can use to automatically obtain dialog act labels for ODP classification.

7 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 authors and do not necessarily reflect the views of the sponsor. We thank several anonymous reviewers for their constructive feedback.

References

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.

Christopher P. Diehl, Galileo Namata, and Lise Getoor. 2007. Relationship identification for social network discovery. In *AAAI*, pages 546–552. AAAI Press.

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.

Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. 2002. Gene selection for cancer classification using support vector machines. *Mach. Learn.*, 46:389–422, March.

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.

Thorsten Joachims. 1999. Making Large-Scale SVM Learning Practical. In Bernhard Schölkopf, Christopher J.C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, Cambridge, MA, USA. MIT Press.

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.

William M. O'Barr. 1982. *Linguistic evidence: language, power, and strategy in the courtroom*. Studies on law and social control. Academic Press.

Philip V. Ogren, Philipp G. Wetzler, and Steven Bethard. 2008. ClearTK: A UIMA toolkit for statistical natural language processing. In *Towards Enhanced Interoperability for Large HLT Systems: UIMA for NLP workshop at Language Resources and Evaluation Conference (LREC)*.

John C. Platt. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *ADVANCES IN LARGE MARGIN CLASSIFIERS*, pages 61–74. MIT Press.

Vinodkumar Prabhakaran, Owen Rambow, and Mona Diab. 2012. 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).

Sacks, E Schegloff, and G Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50:696–735.

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