Pranav Anand, Joseph King, **Jordan Boyd-Graber**, Earl Wagner, Craig Martell, Douglas W. Oard, and Philip Resnik. **Believe Me: We Can Do This!**. *The AAAI 2011 workshop on Computational Models of Natural Argument*, 2011, 5 pages.


```
@inproceedings{Anand:King:Boyd-Graber:Wagner:Martell:Oard:Resnik-2011,
Title = {Believe Me: We Can Do This!},
Author = {Pranav Anand and Joseph King and Jordan Boyd-Graber and Earl Wagner and Craig Martell and Dougla
Booktitle = {The AAAI 2011 workshop on Computational Models of Natural Argument},
Year = {2011},
Location = {San Francisco, CA},
Url = {http://cs.umd.edu/~jbg//docs/persuasion.pdf},
}
```


Links:

- Data [https://sites.google.com/site/persuasioncorpora/]
- Presentation [http://cs.umd.edu/~jbg/docs/persuasion_talk.pdf]

Contact Jordan Boyd-Graber (jbg@boydgraber.org) for questions about this paper.

Believe Me—We Can Do This! Annotating Persuasive Acts in Blog Text

Pranav Anand
UC Santa Cruz
panand@ucsc.edu

Joseph King
UC Santa Cruz
jokking@ucsc.edu

Jordan Boyd-Graber
University of Maryland
jbg@umiacs.umd.edu

Earl Wagner
University of Maryland
ewagner@umiacs.umd.edu

Craig Martell
The Naval Postgraduate School
cmartell@nps.edu

Doug Oard
University of Maryland
oard@umd.edu

Philip Resnik
University of Maryland
resnik@umd.edu

Abstract

This paper describes the development of a corpus of blog posts that are annotated for the presence of attempts to persuade and corresponding tactics employed in persuasive messages. We investigate the feasibility of classifying blog posts as persuasive or non-persuasive on the basis of lexical features in the text and the tactics (as provided by human annotators). Annotated tactics provide substantial assistance in classifying persuasion, particularly tactics indicating formal reasoning, deontic obligation, and discussions of possible outcomes, suggesting that learning to identify tactics may be an excellent first step to detecting attempts to persuade.

1 Introduction

Following Austin's well-known typology of speech acts, communication is often regarded at two levels, what an utterance literally means (the 'locutionary act') and what is meant in context (the 'illocutionary function') (Austin 1962). Much of contemporary natural language processing work in semantics is located somewhere in this territory, be it in terms of determining entailment patterns, sentiment analysis, or characterizing indirect speech acts. But Austin also drew attention to speech acts operating at a 'perlocutionary' level, such as flattering, insulting, and scaring. These are characterized not in terms of the information the utterer was conveying, but the psychological effect on the listener – not the what of an utterance, but the why.

This paper reports our initial step toward computational detection of perlocutionary speech acts. We concentrate on the act of **persuasion**: instances where an agent attempts to convince another party to adopt a novel belief, attitude, or commitment to act. We develop a corpus of over 4,600 blog posts[1] annotated for the presence of persuasion and *tactics* that persuasion theoreticians have argued accompany persuasion attempts, and we present preliminary systems for detecting acts of persuasion using oracle-generated tactics.

Persuasion is one of the most widely studied perlocutionary acts, with links to philosophy (Searle 1969), rhetorical structure (Marcu 1997), and argument modelling (Walton, Reed, and Macagno 2008); it also commands a lively social science literature (Cialdini 2000). Persuasion identification is potentially applicable to broader analyses of interaction, such as the discovery of those who shape opinion or the cohesiveness and/or openness of a social group.

This paper is organized as follows. Section 2 provides an overview of persuasion from the communication sciences perspective, grounded in terms of specific tactics that have been implicated in planning a persuasive communication. In Section 3, we present these tactics in more detail, and Section 4 describes the annotations of a moderately large collection of blog posts both for tactics and persuasion. Section 5 then presents initial results of a feasibility study showing that persuasion tactics are substantially better features than lexical and topic features for identifying persuasive acts.

2 Persuasion and Persuasive Acts

At its most general, persuasion describes when one party (the 'persuader') induces a particular kind of mental state in another party (the 'persuadee'). Thus, like flattery or scaring, but unlike expressions of sentiment, persuasion includes the potential change in the mental state of the other party. Contemporary psychology and communication science further require the persuader to be acting intentionally. Correspondingly, any instance of (successful) persuasion is composed of two events: (a) an attempt by the persuader, which we term the *persuasive act*, and (b) subsequent *uptake* by the persuadee. In this paper, we focus solely on persuasive acts, leaving the question of uptake for future work.

Three primary intended mental state types are recognized in the psychological and communication science literature on persuasion (Miller 1980):

- **Belief Revision:** Acceptance of the truth or falsity of a proposition.
- **Attitude Change:** Adoption of a category of judgment toward some object (an entity, event, or proposition).
- **Compliance Gaining:** Commitment toward or against a course of action.

In principle, the means employed by a persuader (what we term tactics) depend upon the type of intended outcome. For example, a persuasive act involving the proposition that smoking causes cancer (an instance of belief revision) may involve appeals to evidence from outside experts, while a

[1]The annotated corpus and the complete annotation instructions are available at http://sites.google.com/site/persuasioncorpora/.

persuasive act seeking to induce the persuadee to quit smoking (an instance of compliance gaining) may involve reminding the persuadee of the benefits she will accrue if she complies. The different tactics, again in principle, may lead to corresponding differences in linguistic patterns, and hence different detection strategies.

The essence of a persuasive act is the manner by which a persuader attempts to influence the persuadee (i.e., the rhetorical tactics she chooses to trigger successful uptake). Among the Austinian perlocutionary acts the use of tactics is particular to persuasion; flattery acts or scaring acts do not make use of a carefully constructed plan of attack (Marcu 1997). Because they intend to alter the persuadee's mental state, persuasive acts assume that the persuadee is possibly resistant, and the tactics acknowledge this possibility.

The importance of tactics as markers of this potential skepticism (and thereby markers of persuasive acts) can be seen in the contrast between the blog extracts in (i-iii). Neither the ranting narrative in (i) nor the catalog of opinions in (ii) are persuasive acts, as evidenced by the lack of any attempt to engage with a disagreeing reader. In (iii), however, the author attempts to ground an argument that altruism does not exist by making *social generalizations* about how people habitually behave, thereby preemptively responding to someone who doubts the main thesis.

(i) So much for Texas. I almost made it to Dallas tonight, but unfortunately weather and air traffic conspired against my trip. I read most of a novel (Nobody's Fool by Richard Russo if you're keeping tabs) and ate the worst food ever. The lettuce on my club sandwich from TGIFridays was so rotten I had to WIPE it off my nasty meats. I couldn't pick it off. It was that far gone.

(ii) Ok, some quick suggestions and observations . . . Go see Avenue Q. If you like Turkish food and you're in the city, try Sip Sak, but don't order the Lamb & Okra. Get a salad or something grilled. What was I thinking? Okra? Gap clothes fit better than Banana Rep and Express this fall, for all you metros out there.

(iii) Altruism is an illusion. We are all consumers, operating in our own self-interests and the interests of those like us. Without the constructs of "good" and "evil" we will have a better perspective to interpret the media's representation of our world. I am not willing to give up the luxuries and conveniences that we Americans consider unalienable rights more than anyone else.

3 Tactics of Persuasion

We developed a classification of tactics from two prominent tactics ontologies in the social sciences and from a preliminary examination of the blogs described in Section 4. Cialdini (2000) details six non-logical 'principles of influence,' including people's respect for: popular opinion, people who have done them favors, and the suggestions of those they admire or esteem. Marwell and Schmitt (1967) provide twelve strategy types for securing behavioral compliance ('compliance gaining tactics'), involving a mixture of promises/threats, appeals to self-image, and altering the persuadee's temperament. Combining overlapping tactics resulted in fourteen types of tactics. In Table 1 we group these into four basic categories (postulating potential OUTCOMES of uptake, arguments based on types of GENERALIZATIONS, appeals to EXTERNAL AUTHORITIES, and INTERPERSONAL tactics).

Additionally, our examination of the data led us to add the fifth group in Table 1. In a preliminary examination of the blogs described Section 4, we found instances of persuasion without tactics from the first four groups in Table 1, but containing language associated with additional patterns of argumentation, such as some of the argumentation schemes found in (Walton, Reed, and Macagno 2008). In a small set of cases, these were identified as instances of analogy or metaphor designed to frame an issue in a different light (thus, cases of arguments from analogy or definition):

Like the south and slavery, religion is a way of life.

More common were instances involving arguments from causal reasoning ("because", "so that"), arguments from absurdity (i), and arguments from example (ii):

(i) soo, by that logic, 'only 500' would be quite acceptable as an argument too. Ridiculous.

(ii) Pandering to Islamic terrorism has only ever resulted in more of it. Case in point: The Phillipines, where a long-dormant Islamic terrorist outfit, revitalised by the Phillipines' government's cowing to to terrorist demands and pulling troops out of Iraq to free a single hostage, has probably doomed hundreds, if not thousands, to death.

To reduce the strain on annotators, we instructed them to label any argumentative pattern that was not an instance of Redefinition with "Reason," a general purpose tag (an issue we return to in the conclusion).

4 Annotation

We elected to annotate blogs selected from the Blog Authorship Corpus (Koppel et al. 2006), which contains posts from 19,320 different blogs that were gathered from Blogger.com during August 2004. Blogs are a convenient choice for this initial research, since they are easy to obtain and and there is likely to be practical value in detecting attempts to change belief or attitude using blog posts. This specific collection was chosen because of its broad coverage of topics, its monologic structure (as opposed to chat transcripts), its clear authorship (as opposed to speeches and sermons, which may be written by committee), its broad range of register from formal to extremely informal, and because it raises no intellectual property and privacy issues.

For the purposes of determining annotation guidelines and annotator training, 30 posts (distinct from the annotated set) were hand-selected from the larger Blog Authorship Corpus based on their coverage of tactics and persuasion acts. A pilot annotation by the authors with limited guidance yielded low interannotator agreement ($\kappa = 0.4$), principally due to three factors: (a) confusion between expressions of opinion and persuasion, (b) instances where it was not obvious whether an author was intending to persuade, and (c) difficulty in distinguish between belief revision and attitude change (Is "smoking is bad for your health" aimed at attitude change or belief revision? What about "smoking is dangerous"?).

Tactics by Category	Sources
Outcomes	
Threat/Promise. Poses a direct threat or promise to the persuadee.	C
Social Esteem. States that people the persuadee values will think more highly of them.	MS
Self-Feeling. States that uptake will result in a better self-valuation by the persuadee.	C
Outcome. Mentions some particular consequences from uptake or failure to uptake.	MS, C
Generalizations	
Deontic Appeal. Mentions duties or obligations.	MS
Moral Appeal. Mentions moral goodness, badness, etc.	MS
Social Generalization. Makes generalizations about how some particular class of people tendentially behaves.	MS, C
Good/Bad Traits. Associates the intended mental state with a "good" or "bad" person's traits.	C
External	
Popularity. Invokes popular opinion as support for uptake.	MS, C
VIP. Appeals to authority (bosses, experts, trend-setters).	C
Interpersonal	
Favors/Debts. Mentions returning a favor or injury.	C
Consistency. Mentions keeping promises or commitments.	C
Empathy. Attempts to make the persuadee connect with someone else's emotional perspective.	C
Scarcity. Mentions rarity, urgency, or opportunity of some outcome.	C
Other	
Redefinition. Reframes an issue by analogy or metaphor.	W
Reason. Provides a justification for an argumentative point based upon additional argumentation schemes e.g., causal reasoning, arguments from absurdity.	W

Table 1: Common rhetorical tactics for persuasive acts contributed by Marwell and Schmitt (MS), Cialdini (C), as well as argumentative patterns inspired by Walton et al (W).

Based on the trial annotations, guidelines were constructed emphasizing the importance of justificatory text for persuasive language as well as a focus on *blatant* persuasion (cases where an author makes clear her persuasive intention). An analogous procedure was followed for each of the 13 tactics identified above (these instructions are available with the corpus; see fn. 1). Annotators were instructed to mark the smallest text span containing a tactic and to not assume that the presence of a tactic necessarily signaled persuasion (the key arbiter being whether that particular pattern was intended to forestall potential skepticism in the reader). Eight annotators were then trained on a subset of this 30 blog post collection and tested on the remainder, resulting in $\kappa > 0.8$ on the training material. The original set of 19,320 blogs were then revisited and, among those containing more than 200 posts, 40 were randomly selected. These 40 blogs contained 25,048 posts. Each annotator annotated seven blogs (i.e. all selected posts from each blog), with 20% overlap across annotators. Of the 25,048 posts read from 40 blogs, 4,603 posts in 37 blogs were found to contain either persuasive acts or persuasion tactics.

Although a persuasive act may occur over several posts, we annotated persuasive acts at the post level because blog posts are often written as self-contained units. We asked annotators to identify the presence or absence of a persuasive act, and, when persuasion was present, whether belief revision or compliance gaining was the goal. The number of posts containing persuasive acts was small—only 457 of the $25,048$ posts were annotated with any type of persuasion. Of these, 380 were identified as belief revision and 128 were annotated as compliance gaining (51 had both). Because of the relative sparsity of annotation by each annotator, inter-annotator agreement for the corpus was calculated using Krippendorff's α. Overall agreement on persuasive acts was quite reasonable at $\alpha = 0.84$. In view of the relative sparsity of compliance gaining, we merged all types of persuasive acts in the experiments below (the finer annotations, however, are in our annotated corpus). As shown in Table 2, agreement on the tactics themselves was mixed, correlating roughly with their association with particular lexical items (Good/Bad Traits and Deontic Appeal vs. Empathy). For Redefinition, disagreements were contentful: is calling someone e.g., a criminal metaphorical or a value judgment?

5 Predicting Persuasion from Tactics

A guiding assumption of our annotation was that persuasive blog posts could be more reliably detected by systems that take into account tactics. To test this, we conducted a preliminary feasibility study compared the performance of systems that classified posts based upon features extracted from the text with those using human-annotated tactics. Such 'oracle' studies are useful for understanding the potential performance gain of a particular class of features, assuming perfect detection of that class. We evaluated systems with three standard classification results: Precision (the percentage of posts the system called persuasive that were in fact persuasive), Recall (the percentage of persuasive posts that the system classified as persuasive), and F-score (the harmonic mean of

Tactic	Freq	α
Reason	408	0.76
Deontic Appeal	154	0.85
Popularity	114	0.80
Redefinition	109	0.60
Empathy	94	0.71
Outcome	76	0.70
Impt Person	57	0.63
Favors/Debts	55	0.72
Consistency	53	0.84
Good/Bad Traits	31	0.89
Scarcity	11	0.40

Table 2: Persuasion tactics, showing number of passages annotated by the primary annotator with the tactic (Freq) and inter-annotator agreement (Krippendorff's α)

Precision and Recall), which prevents extremes in Recall or Precision from inflating performance.

For our starting baseline system, we considered stemmed unigram counts alone (i.e., using only words with morphology removed). An SVM^{light} classifier (Joachims 1999) was trained on the 4,603 posts using a linear kernel and tested via leave-one-out cross-validation. The resulting classifier was reasonably precise, but suffered from low recall (Precision=0.742, Recall=0.174, F-Score=0.282).

The poor performance of the baseline lexical system suggested that less word-dependent features might prove helpful. We considered three sources of more general categories: a) the presence of word classes dealing with sentiment, causation, and insight; b) the presence of particular topics; and c) the presence of tactics. For (a), we extracted 71 **List** count features from the Linguistic Inquiry and Word Count (Pennebaker and Francis 1999) and MPQA subjectivity lexicon (Wilson 2008), two dictionaries categorizing words into more general semantic classes. For (b), we chose to model topic in terms of the generative Latent Dirchlet Allocation model (Blei, Ng, and Jordan 2003), which views each topic as a probability distribution over all words in the corpus; correspondingly, each post may be seen as a probability distribution over topics. 25 "topics" were computed over the 4,603 posts with a symmetric Dirichlet prior, and each post was assigned 25 **LDA** features, corresponding to the amount of each topic in the post. Finally, to test the utility of tactics, we provided systems with 14 oracle **Tactic** count features, derived from the human-annotated tactic labels. To assess the degree to which these feature sets complement each other, we built a Naïve Bayes classifier for each of the 7 possible combinations of them, a method known as an *ablation study*. Table 3 summarizes the results for each combination. Tactic features alone produce substantially better results than any other combinations of features.

To determine the relative importance of each tactic to the classification results, we performed a further ablation study over the 14 tactic features. We found that Reason labels were the primary contributors to classifier accuracy, followed by Deontic Appeal and Outcome. Four additional features were

	Naïve Bayes		
feature set	P	R	F
Tactic	0.505	0.677	0.579
Tactic+LDA	0.161	0.401	0.229
Tactic+List	0.093	0.170	0.121
Tactic+LDA+List	0.109	0.228	0.147
LDA	0.114	0.271	0.161
LDA+List	0.099	0.205	0.133
List	0.079	0.141	0.101
SVM Baseline	0.742	0.174	0.282

Table 3: Relative Effectiveness of Feature Classes in terms of **P**recision, **R**ecall, and **F**-score

	Naïve Bayes		
feature set	P	R	F
DRO+EReThTr	0.509	0.674	0.580
DR+EReThTr	0.510	0.631	0.564
DO+EReThTr	0.456	0.326	0.38
RO+EReThTr	0.513	0.634	0.567
D+EReThTr	0.421	0.254	0.317
R+EReThTr	0.510	0.576	0.541
O+EReThTr	0.418	0.233	0.299
DRO	0.514	0.643	0.571
All except DRO	0.327	0.233	0.272

Table 4: Relative Effectiveness of Tactic Features [**D**eontic Appeal, **E**mpathy, **O**utcome, **R**eason, **Re**characterization, Good/Bad **Tr**aits, **Th**reat/Promise]

responsible for slight gains (Empathy, Recharacterization, Threat/Promise, and Good/Bad Traits); Table 4 shows the results for Naïve Bayes. For deeper understanding, we trained the rule-based classifier RIPPER (Cohen 1995) over these seven features. The classifier learned rules where the first three are consistently positively correlated with persuasive acts, as expected. Surprisingly, the latter four were negatively correlated, all in conjunction with Reason. In manual inspection, 65% of these instances were narratives, where linguistic devices associated with causal reasoning were actually being used as overt markers of discourse relations that structured the narrative (e.g., explaining a course of action) and not used to persuade the reader. These results suggest that while a large number of tactics (e.g., Consistency, Popularity) are not discriminative enough to be reliable markers of persuasion, some are actually useful negative features.

Error Analysis

Given the overall rarity of persuasion in our corpus, we were most interested in posts containing a persuasive act classified as non-persuasive. In the error analysis of such false negatives, we discovered that 2.1% were incorrectly annotated. A larger number were labeled inconsistently because of the strong overlap with sentiment. 16.0% were highly charged, opinionated statements or personal statements that if per-

suasive, have very weak justifications, and do not pass our standard of blatant persuasion (Section 4). As many of these were judged persuasive by 3 annotators, we suspect that the overall class imbalance for not persuasive led annotators to become more liberal of what passed the threshold for persuasive text. 4.0% were reviews of movies, books, and CDs, which were not covered in the guidelines.

68% of the incorrectly labeled posts were legitimately persuasive on reexamination. Of these posts, approximately 10% contained generic statements (e.g., "Single parent kids are always sure - somewhere - that they aren't wanted"), which occur in 2% of the corpus. An additional 3% of the errors contained imperative sentences with the verbs *remember*, *think*, and *imagine*. In aggregate, these comprise 14% of false negatives. Capturing these idiosyncratic patterns requires syntactic knowledge missing from current features, which treat posts as bags of words.

6 Conclusion

In this paper, we tentatively explored the computational detection of authorial intent, focusing on the perlocutionary act of persuasion. Drawing on the rich theory of persuasion and persuasive tactics found in the social sciences, we constructed a system for persuasive act labeling. We have demonstrated that oracle models of tactics, especially statements of logical reasoning, improve classification over word-based, topic-based, and word class baselines systems. We leave building such models to future work, noting that, like for persuasion itself, a unigram SVM baseline for Reason is quite poor (Precision=0.575, Recall=0.103). However, rhetorical relations are often unmentioned in text, and their discovery thus requires richer features. We also would like to revisit the separation of the three types of persuasion.

Given the relative prominence of our general purpose Reason tactic, we would like to consider the argumentative structures exploited in blogs more closely. Many of the tactics in Table 1 are represented in the literature on argumentation modeling (arguments from popularity, from authority, and from threat being classic examples from rhetorical theory). Interestingly, only a few were sufficiently prevalent in our blog search. Revisiting the Reason category will allow us to consider other argumentative schemes that have received less attention in the psychological study of persuasion, such as arguments from contradiction or analogy.

Finally, the annotated instances of persuasion and the rough metrics we can learn from the tactics will prove useful in more intelligent expansion of the corpus, by targeting particular authors and linguistic features which are more likely to be persuasive. While it might seem surprising on first blush that so little persuasion was found in our initial annotation, for many bloggers persuasion is a distant third aim, behind cathartic release and narrative update. In this regard, our corpus is likely to have different features from those explored in recent attempts to annotate and classify argumentative sections from legal texts in the European Court of Human Rights and Auracania (Mochales and Ieven 2009; Palau and Moens 2009). For instance, we contended with differentiating statements of opinion from attempts to persuade; it is unclear if such an issue arises in more clearly argumentative genres.

The rise of social media has rendered it possible to track the dynamics of the spread of ideas, as well as those who make that spread happen. As this work grows in importance, we believe that detecting persuasion will be a key component of many tasks of interest. We hope that providing our annotated resource of persuasion and persuasive tactics will allow other researchers to build tools to better understand these phenomena.

Acknowledgments

This work was funded by the Intelligence Advanced Research Projects Activity (IARPA) through the Army Research Laboratory.

References

Austin, J. L. 1962. *How to do things with words*. Cambridge, Mass.: Clarendon.

Blei, D. M.; Ng, A.; and Jordan, M. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* 3:993–1022.

Cialdini, R. B. 2000. *Influence: Science and Practice (4th Edition)*. Allyn & Bacon.

Cohen, W. 1995. Fast effective rule induction. In *Proceedings of the 12th International Conference on Machine Learning*, 115–123.

Joachims, T. 1999. Making large-scale SVM learning practical. In Schölkopf, B.; Burges, C.; and Smola, A., eds., *Advances in Kernel Methods - Support Vector Learning*. Cambridge, MA: MIT Press. chapter 11, 169–184.

Koppel, M.; Schler, J.; Argamon, S.; and Pennebaker, J. 2006. Effects of age and gender on blogging. In *In AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs*.

Marcu, D. 1997. Perlocutions: The achilles' heel of speech act theory. *Journal of Pragmatics*.

Marwell, G., and Schmitt, D. 1967. Dimensions of compliance-gaining behavior: An empirical analysis. *sociomety* 30:350–364.

Miller, G. R. 1980. *The Persuasion Handbook: Developments in Theory and Practic*. Beverly Hills, CA: Sage. chapter On being persuaded: Some basic distinctions.

Mochales, R., and Ieven, A. 2009. Creating an argumentation corpus: do theories apply to real arguments? A case study on the legal argumentation of the ECHR. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Law*, 21–30.

Palau, R., and Moens, M. 2009. Argumentation Mining: The Detection, Classification and Structure of Arguments in Text. In *Proceedings of The Twelfth International Conference on Artificial Intelligence and Law*, 98–107.

Pennebaker, J. W., and Francis, M. E. 1999. *Linguistic Inquiry and Word Count*. Lawrence Erlbaum, 1 edition.

Searle, J. R. 1969. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge, London: Cambridge University Press.

Walton, D.; Reed, C.; and Macagno, F. 2008. *Argumentation Schemes*. Cambridge University Press.

Wilson, T. A. 2008. *Fine-grained Subjectivity and Sentiment Analysis: Recognizing the Intensity, Polarity and Attitudes of Private States*. Ph.D. Dissertation, University of Pittsburgh.