An Exploration of Text Analysis Methods to Identify Social Deliberative Skill[1]

Tom Murray, Xiaoxi Xu, Beverly Park Woolf

School of Computer Science, Commination Dept.
University of Massachusetts, Amherst, MA

Contact: tmurray@cs.umass.edu

Abstract: We report on text processing and machine learning methods with the goal of building classifiers for social deliberative skill, i.e. the capacity to deal productively with heterogeneous goals, values, or perspectives. Our corpus includes online deliberative dialogue from three diverse domain contexts. We use the LIWC and CohMetrix linquistic analysis tools to generate feature sets for machine learning. We report on our evaluation of various machine learning algorythms, feature selection methods, and cross-domain training methods.

1 Introduction

A key human capacity is the ability to negotiate situations involving differing opinions where a resolution of ideas is sought, e.g., in dispute resolution, collaborative problem solving, bargaining, and civic deliberation processes. The need for this deliberative capacity, which we call social deliberative skill (SD-skill), is seen in all realms of human activity from international politics, to collaborative work, to mundane familial squabbles. As communication, collaboration, and deliberation occur increasingly on the internet we believe that there is great potential to design software that supports skillful deliberation through gentle prompts and scaffolds, especially for groups of interlocutors who, acknowledging that deliberation in complex and stressful situations can be challenging, are interested in putting some attention and effort on the quality of their communication. Our overall research goals are to better *understand, assess, and support* SD-skills in online contexts. Our evaluation of software features designed to support SD-skills is reported elsewhere (Stephens, et al. 2103 in submission). Evaluation of SD-skills in that study used a hand-coding scheme. Here we focus on our attempts to use machine learning to assess or model SD-skills based on participant text. Automated assessment will not only facilitate *data analysis* by allowing us to assess more data faster, but, if done in real time, can be used in visualization tools for SD-skills and other important dialogue and deliberation metrics. We have prototyped a Facilitators Dashboard tool that gives facilitators, teachers, or par-

[1] An extended version of this paper is available on the first author's web site.

adfa, p. 1, 2011.

ticipants a birds-eye view of conversation metrics, as described in (Murray et al., 2013, in submission).

2 Background

Social deliberative skills. We frame SD-skills in terms of these capacities (see Murray et al., 2013 submitted): perspective taking (includes cognitive empathy, reciprocal role taking); perspective seeking (includes social inquiry, question asking skills); perspective monitoring (includes self-reflection, meta-dialogue); and perspective weighing (related to "reflective reasoning" and includes comparing and contrasting the available views, including those of participants and external sources and experts). SD-skills overlaps with but is distinct from other cognitive constructs that have been studied in depth, including collaboration skills, metacognition, reflective reasoning, social intelligence, argumentation skills, and critical thinking . We differentiate our research from others that focus on *argumentation*, which aims to help learners generate logical, well-formed, well-supported explanations and justifications (Andriessen et al., 2003), usually framed in objective rather than intersubjective terms. That is, they are about finding the right answer or the most efficient and effective solution to a technical or scientific question—but don't address, as we do, the *skills need in those moments during deliberation or collaboration containing opportunities for mutual understanding and mutual recognition*.

Text Classification. Text analysis has been used successfully for a wide variety of purposes, including to: grade essays (Shermis & Burstein 2003), analyze content for conceptual understanding (Lintean et al., 2011), score text sophistication, writing quality, and reading grade level (McNamara et al., 2010), and score deliberative, argumentative, and question-answering quality (Rose et al. 2008; Ravi & Kim 2007). Past research exploring linguistic and discourse features in dialogues has proven moderately successful in predicting complex phenomena such as personality type, status, deception behavior, metacognition, speech acts, intention, and affect states . Therefore, it is plausible to expect that a linguistic and discourse analysis of deliberation dialogues would provide valuable insights into predictors that are diagnostic of deliberation dynamics and skills. Our research question is whether such methods can be used to predict SD-skills.

Our primary goal is to build domain-independent classifier models that will predict what we call Total-SD-skill, and, later, individual SD-skill components (the total skill is a summation of individual skill occurrences). Perhaps the most prominent machine learning method used in natural language processing, information retrieval, and document/text classification is the 'bag of words" unigram method, in which the feature set for the learning algorithm consist of an unordered set of all the words in a document (preprocessed with stemming etc. as necessary). However, we have much more information available with which to build our predictive models, including deep and surface text classification metrics previously researched. In particular, CohMetrix (Graesser et al., 2011) and LIWC (Linguistic Inquiry Word Count; Pennebaker et al., 2007) are two highly cited and used text analysis in systems in domains related to dialogue and collaborative learning. We hypothesize that using LIWC and CohMetrix outputs as feature inputs to machine learning models would increase their accuracy and efficiency vs bag-of-words methods. Thus we can do a two-step analysis, in which we extract the CohMetrix and/or LIWC features, and then use these features as inputs to machine learning methods.

3 Method and Results

Coding: We have developed and refined a 30-category hierarchical coding scheme for human raters to code segments of the text according to speech act type (which, for our purposes, is sometimes equivalent to SD-skill indicators) showing inter-rater Cohen's Kappa statistics of 71% on average in these domains (Murray et al, 2012). For this paper we focus on a Total-SD-skill metric that is true if any of 17 codes associated with higher quality deliberation is true (including: perspective taking, asking clarifying questions, mediation actions, and meaning generation and repair actions, weighing alternatives, citing sources, changing ones mind, and apologizing). **Corpora**: Table 1 shows descriptive statistics for the three domains we have coded, civic deliberation postings from a neighborhood civic engagement online discussion forum; email exchanges from a faculty listserv where two research communities were engaged in a negotiation discussion; and postings from 7 online discussions on controversial issues from three college classrooms.

Domain	Posts	Seg-ment	Partic-ipants	SD-Skill seg	% SD-Skills	Words/ Post	Posts / Partic	Seg. / post
Civic deliberation	51	396	31	225	57%	352	1.6	7.8
Faculty negotiation	72	438	16	231	53%	195	4.5	6.1
College discussions	768	1783	90	565	32%	88	8.5	2.3

Table 1: Descriptive Statistics for Three Domains

Results. Results can only be sketched in this short paper. Early work looked at correlations between LIWC and CohMetrix measurements and the individual and Total-SD-Skill manual classifications. There were a number of small correlations, such as LIWC "Assent" 8.5% (R-squared) with AGREE speech acts; and CohMetrix SecondPersonPronoun 4.4% with INTERSUBJECTIVE speech acts. The top 20 correlations were in the 1% to 4% range. Though there was not obvious strong correlation between individual LIWC/CohMetrix measure and manual codes, there were a number of smaller correlations that indicated that a machine learning algorithm might combine these to predict the codes.

In our first attempts at building a model for Total-SD-skill we used standard SVM (support vector machine) methods and found that none of the models using LIWC and CohMetrix measurements did as well as the unigram bag-of-words features (we tried using the full set of LIWC and CohMetrix measures and a subset of measures highly correlated with Total-SD-skill). (Note that in this document we used 10-fold cross validation where applicable on all machine-learning methods, unless otherwise stated; SVM used unigram features TF-IDF settings). As expected, we found that trying to predict individual SD-skills was much more challenging than predicting Total-SD-skill, so we focused on Total-SD-skill. Next we compared several machine learning methods: SVM (Cortes & Vapnik, 1995), Naïve Bayes (Rish, 2001), and L1 Regularized Logistic Regression (Tibshirani, 1996) (trying various tuning parameters for each to arrive at a best-guess parameter set). The best performance was obtained using the LI RLR method using the LIWC and CohMetrix measures as features. L1 RLR is purported to have superior generalizability, interpretability, and scalability vs. other methods.

Next we turn to the question of whether some deliberative domains make better training sets for a domain-independent model (see Xu et al., 2013 submitted). We hypothesized that domains that have least skew (imbalanced frequency distributions) might serve as better training sets. Results include: (1) Overall using the **Civic domain** as the training set did much better than using the Faculty domain, the Classroom domain, or all of the data as the training set. This was true for all three learning algorithms and all four performance measures (accuracy, precision, recall, and F2). Our hypothesis that the domain with the least skew would serve as the best cross-domain training set was confirmed. (2) Overall the **L1 RLR algorithm** significantly outperformed Naïve Bayes and SVM (this was true when the Civic or Faculty domains were used to train). This confirms our expectation that L1 RLR has performance characteristics addressing for the modeling challenges we face. (3) From #1 and #2 above we see that the **best model for domain-independent** prediction, i.e. prediction that worked best averaged over all three domains, was L1 RLR using the Civic domain for training: accuracy 51%, precision 49%, recall 82%, and F2 71%. (4) ***Cross-training*** proved to have advantages. For precision, recall, and F2-measure (but not accuracy) using the Civic domain as a training set outperformed using the *same* domain to train as was tested on. I.E. for performance on the Faculty domain by itself, training with Civic was better than training with Faculty. Similarly with the Classroom domain. (5) These overall results for binary classification of Total-SD-skill, accuracy 51%, precision 49%, recall 82%, and **F2 71%**, are encouraging for our exploratory study, but not particularly impressive for a binary classifier.

References

1. Graesser, A.C., Dowell, N., & Moldovan, C. (2011). A computer's understanding of literature. Scientific Study of Literature 1:2 (2011).
2. Ravi, S., Kim, J. (2007). *Profiling student interactions in threaded discussions with speech act classifiers*. IOS Press.
3. Lin, X. & Sullivan, F. (2008). Computer contexts for supporting metacognitive learning. In J. Voogt, G. Knezek (eds.) International Handbook of Information Technology in Primary and Secondary Education, 281–298. Springer Science+Business Media, LLC.
4. McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic features of writing quality. Written Communication, 27(1), 57-86.
5. Murray, T., Wing, L., Woolf, B., Wise, A., Wu, S., Clark, L. & Osterweil, L. (2013 in submission). A Prototype Facilitators Dashboard: Assessing and visualizing dialogue quality in online deliberations for education and work. Submitted to Elearn-2013.