Quantitative Estimation of the Strength of Agreements in Goal-Oriented Meetings

Been Kim, Larry A.M. Bush and Julie Shah
Massachusetts Institute of Technology
Cambridge, Masachusetts 02139
{beenkim, BushL2, julie_a_shah}@csail.mit.edu

***Abstract*—Ineffective meetings occur frequently and participants leave with different understandings of what has been decided upon. For meetings that require quick responses (e.g., disaster-response planning), everyone must leave the meeting on the same page to ensure the successful execution of the mission. Detecting patterns of weak agreements in planning meetings is the first step towards designing an intelligent agent that encourages team members to revisit decisions that may adversely affect the team's performance, and to spur dialog that results in higher quality plans. This paper presents a statistical approach to learning patterns of strong and weak agreements without using domain-specific content or keywords, meaning the algorithm takes as input information about how the team plans but does not require potentially sensitive data on what is being planned. Our approach applies statistical machine learning to *dialog features*, which prior studies in cognitive psychology have shown qualitatively capture the level of joint commitment to plan choices. We analyze a real-world conversation dataset, the AMI corpus, to quantitatively verify that *dialog features* improve the estimation of strength of agreements over prior approaches. We show these results are consistent across a number of different supervised and unsupervised learning algorithms, and that can achieve up to 94% average accuracy in estimating the strength of agreements.**

I. Introduction

Goal-oriented meetings are frequent occurrences in our everyday lives. For example, we discuss project plans at work and coordinate with friends to organize outings. The consequences are inconvenient but relatively minor if the participants leave these types of discussions with different understandings of what was decided upon. Yet, in high-intensity domains such as disaster-response, minor differences in understandings may degrade the team's ability to successfully coordinate and may have serious consequences for people's safety.

It is challenging for a group to reach consensus through dialog. The interaction among people is complicated, dynamic and unpredictable. Natural human collaborative dialog unfolds in cycles, agreements are fluid, and proposals are often implicitly communicated and accepted [7]. In addition, there are social aspects that are often hard to formulate into equations. These characteristics make explicit modeling and quantitative analysis of goal-oriented dialog challenging.

Prior art provides a theoretical foundation for translating the ambiguous and inconsistent nature of dialog into a set of *dialog features* that indicate that level of joint commitment to plan decisions [7]. In this paper, we generalize this qualitative approach to enable a quantitative, predictive capability for characterizing weak and strong agreement in dialog. We envision this work as a first step towards the design of an intelligent agent that observes human team planning and interjects to highlight weak agreement among team members. This approach would integrate an intelligence agent into natural human team dynamics and does not require extensive codifying of domain knowledge. In contrast, prior approaches to decision support in planning utilize automated planners to provide suggestions to the team. A common criticism of this approach is that automated planners cannot practically capture all relevant domain knowledge and expertise and too frequently provide uninformative solutions [25].

Our approach to estimating the strength of agreements builds on prior, qualitative investigations of the human team decision-making process ([7], [1], [16]). Traditional approaches use dialog acts (DAs) to capture the role of an utterance (e.g., suggestion, information-request). DAs are widely used in natural language processing systems, for example, to predict the next dialog [22] and identify agreement and disagreement [10], [13], [2], [15]. It is also possible to automatically tag DAs as shown in [27], [18].

However, recent empirical studies (e.g., Eugenio et al. [7]) have shown that the traditional DAs (such as accept and reject) cannot capture the level of joint commitment and only consider one agent's attitude toward an action. Eugenio et al. propose a new set of dialog features (referred to here as Eugenio's features) that capture the level of joint commitment as the negotiation unfolds and demonstrate that the features are very important to accurately estimating the strength of agreements. Eugenio's features track how commitment evolves from an inability to commit (*partner decidable option*), to conditional commitment (*proposal*), to *unconditional commitment*. In this work, we aim to translate qualitative studies in cognitive psychology into a quantitative predictive framework and confirm that dialog features play an important role in estimating the strength of agreements.

One of the key benefits in using dialog acts and features to characterize the strength of agreements is that this approach does not require the extraction of keywords or other content information from the dialog. In other words, this approach uses information about how the team plans, but does not require storing and processing potentially sensitive information about

what they are planning as is the case for previous quantitative approaches to this problem ([13], [15]). To our knowledge, our approach is the first to (1) estimate the strength of agreements based solely on DAs and features and not keywords, and to (2) map the qualitative theoretical foundations ([7]) for strength of agreements to a quantitative, predictive measure.

In this paper, we conduct statistical analysis of human meetings for a publicly available meeting dataset ([21]). First, we validate the use of Eugenio's features for estimating the strength of agreements and show improvement over approaches that use traditional DAs alone. Second, we show that the improvement in estimation is consistent throughout different statistical approaches. Third, we quantitatively measure the usefulness of Eugenio's features, compared to the traditional DAs, using two machine learning feature ranking algorithms. We also show that Eugenio's features improve the performance of the estimation task and achieve up to 94% average accuracy for the tested data corpus without using contents or keywords.

Section II places our work in the context of prior art, and Section III describes the dataset we analyzed. Section IV presents our statistical approach to estimate the strength of agreements. Results are presented in Section V, followed by our contributions and future work in Section VI.

II. Related Work

This section briefly summarizes both qualitative and quantitative prior art related to our work.

Qualitative study of the group decision making process is an active area of study. A number of studies focus on designing descriptive models for the group decision making process, such as modeling and understanding the course of agreement [1], developing theories for group decision making [4] and studying different group decision-making schemes [11]. Among others, Hiltz et al. [16] take a qualitative approach for measuring the level of agreement as an evaluation metric to compare face-to-face communication and computerized conferencing. The level of agreement is measured by listening to or looking at the final opinion expressed by each individual in a group, and observing if they are the same.

Prior art in quantitative works to measure the strength of agreements use classification techniques to classify agreement and disagreement given various content information of the meeting dialog. Galley et al. [10] perform classification using lexical, durational and structural features of the conversation. By identifying adjacency pairs, and both looking forward and backwards in the discourse, they achieve 86.9% accuracy in binary classification (agreement or disagreement). Hahn et al. [13] and Bousmalis et al. [2] apply learning methods to classify agreement and disagreement using verbal (e.g., keywords) and non-verbal (e.g., head nods) cues. Hillard et al. [15] perform binary classification with 78% accuracy using word-based features (e.g., the number of positive and negative keywords) and prosodic cues (e.g., pause, frequency and duration). All the above work require extraction of potentially sensitive information and what is being planned in the meeting and/or video clips to perform classification.

More recently, Eugenio et al. [7] conduct a qualitative study focused on how negotiations unfold in the design task. This work finds that the notion of commitment is more useful than that of acceptance or rejection (traditional DAs) in order to model the agreement process. A new set of features (explained in more detail in Section III-C) are introduced to trace how commitments change, and capture the joint commitment of a proposal. These features are qualitatively proven to help in recognizing implicit and/or passive acceptance. This begs the question as to whether these features can help to estimate the strength of agreements in a quantitative study, and can be used as an automatic tool for augmenting human team planning.

The main differences between previous qualitative and quantitative works and ours are that: (1) Our approach maps the qualitative theoretical foundation ([7]) for strength of agreements to a quantitative measure and verifies what has only previously been qualitatively verified. (2) Our approach incorporates implicit/passive acceptance of proposals in estimating the strength of agreements by adopting Eugenio's features†[1]. (3) Our approach does not require extraction of keywords or videos (for non-verbal cues), therefore, does not require exposing potentially sensitive information.

III. Hypotheses and Methodology

In this section, we present our hypotheses and approaches to validate them. Next, we describe dataset used in this work. In addition, we provide the definitions of dialog features, introduced in [7], and how they are tagged in our dataset.

A. Hypotheses and approaches to validation

In this section, we present a list of proposed Hypotheses (H) and Approaches (A) to validate them.

H1 Eugenio's features capture the level of joint commitment among decision-makers, and therefore can be useful to improve the estimation of the strength of agreements, better than with the traditional DAs alone. We expect this trend to hold across different statistical methods.
We define features are useful if any of the following is true: 1) features always improve the estimation accuracy when used with traditional DAs than traditional DAs alone (i.e. the performance is always better than without features) 2) features can replace traditional DAs (i.e. the performance is better by only using Eugenio's features) 3) features are always helpful with traditional DAs (i.e. the performance is better or the same).

A1 We use a number of different supervised and unsupervised learning algorithms to investigate the usefulness of Eugenio's features and show the features improve the estimation accuracy for the strength of agreements.

[1]† Eugenio's features are designed to capture implicit and explicit joint commitments. For more information, please refer to [7].

# data	Strength of agreements label
94	Strong
47	Weak
total 141 data points	

TABLE I: Design task dataset

H2 Eugenio's features are more informative compared to traditional DAs in estimating the strength of agreements. We expect this trend to hold across multiple statistical methods.
We define a feature is more informative than others if the rank of the feature is higher than others in a statistical feature ranking algorithm.

A2 We apply two feature ranking algorithms to learn the role of Eugenio's features.

More details on each algorithm can be found in Section IV.

B. The design problem and dataset

We focus on the design problem in this work. There are a number of problem domains where analyzing the strength of agreements is useful (e.g., planning, scheduling). However, as mentioned in [7], the design problem is suitable when studying the strength of agreements, as it primarily involves the negotiation of product features and the constraints between dependent design sub-tasks. For other design scenarios, please refer to [19] [20].

In our study, we use one of the most extensively annotated corpus of meetings data publicly available, originating from the AMI (Augmented multi-party interaction) project ([21]). This dataset contains a large number of annotated meetings (over 12,000 human-labeled annotations). In total, there are 108947 DAs in a total number of 95 meetings, and 26825 adjacency pairs (explained below). This corpus is derived from a series of real meetings, where each meeting has four participants who work as a team to compose a new design for a new remote control. Each participant takes a role, either project manager, marketing expert, industrial designer or interface designer. The participants are given training for their roles at the beginning of the task. Documents used to train participants and annotators are publicly available. Here we provide description of the annotations provided with the corpus that are relevant to our work.

Dialogue Acts: Dialogue acts (DAs) annotation represents the role of an utterance. It is possible to automatically tag DAs as described in [27], [18]. DAs include questions, statements, suggestions, assess (positive, negative or neutral), understanding etc.* [2] One utterance (one turn that one participant takes) is often divided into pieces and tagged with different DAs.

Decision summary: Decision summary includes a summary of decisions made in a meeting and related DAs tagged by annotators (3rd party). Each annotator is given a list of topic descriptions to choose from, but he/she is allowed to create one if none is found.

Participant summary: Participant summary includes a summary of a meeting created by each of the participants. It includes decisions that each participant reports are made in the meeting.

We used 141 number of discussions (1 topic per discussion) from 47 meetings†[3]. Each meeting consists of 2-5 different topics.

C. Eugenio's feature definitions

Eugenio's features suggested in [7] are designed to monitor the level of commitments of participants as the agreement process unfolds. These features include *proposal* (P), *partner decidable options* (PDO), *unendorsed option* (UO) and *unconditional commitment* (UC). These features can be coded using conventional DAs combined with solution sizes. A solution size is defined to be 'determinate' if there is one or more solutions for a set of constrained parameters‡, [4] and 'indeterminate' otherwise. In other words, the solution size for a parameter is indeterminate if not enough information has been exchanged to make a valid (i.e. fully informed) proposal**[5]. The coding scheme of each feature is shown below:

- **Proposal (P)** corresponds to utterances tagged as all of the following: action-directive (DA), offer (DA) and determinate (solution size).
- **Partner decidable options (PDO)** corresponds to utterances tagged as one of the following: 1) open-option (DA) and indeterminate (solution size) or 2) action-directive (DA), offer (DA) and indeterminate (solution size).
- **Unendorsed option (UO)** corresponds to utterances tagged as all of the following: open-option (DA) and determinate (solution size).
- **Unconditional commitment (UC)** corresponds to utterances tagged one of the following: 1) action-directive (DA), commit (DA) and determinate (solution size) or 2) action-directive (DA), commit (DA) and indeterminate (solution size).

D. Eugenio's feature tags generation

In this section, we briefly explain how we tag annotations that are not provided by AMI dataset but are needed to define Eugenio's features. We also describe how the ground truth labels are tagged.

[2]*More details of definition of dialog acts can be found from [21]

[3]† Only 47 meetings out of 95 meetings come with participant summary, which are required to label each decision to be strong or weak. How these labels are generated is explained in Section III-D

[4]‡Parameters mean topics for the team to discuss (e.g., what should a person A do at time T with who?). The details can be found in [7].

[5]**More examples can be found from [7].

1) Solution sizes: The first tags we need are solution sizes. As mentioned in [7], solution size tagging is straightforward. We apply a simple algorithm to automatically tag solution sizes.

(a) Mark all the information request (elicit-inform type of DAs) within a scope of a parameter.
(b) Find the last information request in the scope, and consider it as the end of the balance of the information.
(c) Mark all the utterances before the last information request as indeterminate solution size, and all the ones after as determinate solution size.

We verify this algorithm by manually tagging 24 data points and comparing the results with the algorithm results. The algorithm agrees with manual tags 90% of the time. Then we tagged the rest of the data points using the algorithm.

2) Action-directives: Secondly, action-directives represent all elicit forms of DAs, which are DAs that require actions from hearers. More details on definitions of DAs can be found from the manuals in [21].

3) The strength of agreements labels: Lastly, we need the ground truth for the strength of agreements. To generate the label, we take the approach similar to the one from [16]. The details of the label generation is explained below:

(a) Find the discussion annotation provided by AMI dataset and collect related DAs for each topic.
(b) Find participant summary provided by AMI dataset.
(c) In participant summary, find the section that summarizes the topic in discussion annotation from (a).
(d) Compare each participant's summary about the meeting on that topic.
(e) If all participants mention the same conclusion on the topic, it is a strong agreement.
(f) If at least one participant mentions different conclusion on the topic, it is a weak agreement.

These labels are manually generated by two annotators, and the kappa coefficient was 0.73. Table I shows the distribution of labels in the dataset.

IV. Quantitative approach

We apply a number of unsupervised mixture model clustering and supervised learning algorithms to estimate the strength of agreements. For supervised algorithms, we perform leave-one-out cross-validation by leaving a single data point as a test dataset, and use the remaining observation as the training data. The supervised algorithms are Support Vector Machine (SVM) with Radial Basis Function (RBF) kernel and logistic regression. For the unsupervised algorithms, we (purposely) discard the annotations and apply Expectation Maximization (EM) with Gaussian Mixture Models (GMM) and Kmeans algorithm. In addition, feature ranking algorithms are applied to verify Hypothesis 2. This section summarizes the details and algorithms used in this work. Note that only a subset of algorithms are explained here. A great reference for the rest of algorithms can be found in [14].

	$+$ Class	$-$ Class
Feature i was observed	a	b
Feature i was not observed	c	d

TABLE II: 2×2 contingency table for Fisher's exact test for Feature i

A. Inputs to algorithms: Feature vectors

The feature vector is a row vector where each element represents the number of corresponding DAs or Eugenio's features in the time frame of interest. For example, one data point can represent the first discussion in the second meeting and can have 3 proposals, 2 positive responses and 10 information exchanges. Once we have feature vectors for all data points, the input to algorithms is a matrix where the rows are indexed by data points (feature vectors) and the columns are indexed by features.

Among DAs, we use only a subset that are known to be relevant to the group decision making process [7] (e.g., accept, reject), or are required to define Eugenio's features (i.e. DAs that are required to code P, PDO, UO and UC: action-directive, offer, open option, commit).

B. Support Vector Machine for classification

The support vector machine (SVM) [3] is one of the most well-known supervised classification techniques. It produces a nonlinear boundary with the objective of maximizing the margin between the training points. SVM has been successfully applied to a number of applications in machine vision [23][24], handwriting recognition [26][5], and bioinformatics [9][17]. The nonlinear decision boundary (i.e. a hyperplane) is defined by

$$x : f(x) = x^T \boldsymbol{\lambda} + \lambda_0 = 0 \tag{1}$$

where $\boldsymbol{\lambda}$ is a unit vector. For optimization, we use sequential minimal optimization (with 30,000 maximum iteration) and with different kernels. We will leave more details and properties of general SVM to [14].

C. Feature ranking

This section, we briefly explain two feature ranking algorithms used in this work. These algorithms are used to verify Hypothesis 2 by measuring the role (i.e. rank) of individual features in estimating the strength of agreements.

1) Support Vector Machine with linear kernel: For SVM with linear kernel, the decision boundary parameter $\boldsymbol{\lambda}$ in Equation 1 contains information about how important each of the features are to classify the data (previously introduced in [12]). In binary classification with $+/-$ class, if $|\lambda_i|$ is large, where λ_i represents the value of the $\boldsymbol{\lambda}$ for the feature i, the corresponding feature i is more important than other features with smaller value of $|\lambda_i|$. Large positive entries of $\boldsymbol{\lambda}$ correspond to strong features for the $+$ class, and small negative entries (i.e., large $|\lambda_i|$) in $\boldsymbol{\lambda}$ correspond to strong features for the $-$ class.

		Subsets of features		
		DA only	Eugenio's + DAs	Eugenio's only
Methods	Logistic	92% (26%)	92.4% (26%)	**94.6% (22%)**
	SVM with RBF	75% (43%)	**90% (29%)**	50% (50%)
	Kmeans	55%	**66%**	60%
	EM with GMM	**66%**	65%	65%

TABLE III: The average estimation accuracy of the strength of agreements. The bold texts represent the best average accuracy for each method. Standard deviation is provided for supervised methods.

	Top important features for	
Rank	Weak agreements	Strong agreements
1	**Proposal**	**Unendorsed option**
2	**Partner decidable options**	Reject
3	Offer	Open-option
4	Action-directive	Commit
5		Accept
6		**Unconditional commitment**
7		Info-request

TABLE IV: SVM feature ranking results . The bold texts are Eugenio's features.

2) Fisher's exact test: Fisher's exact test is used to find associations between two categorical variables ([8]). Essentially, it tests the null hypothesis — if the class label and the feature are independent, the feature does not help classification, therefore the feature is not very important. To perform this test, we build 2×2 contingency table for each of the features as shown in Table II. For example, the variable 'a' represents the number of data where Feature i is observed and the data has $+$ class label. Similarly, the variable 'd' represents the number of data where Feature i is not observed and the data has $-$ class label.

We reject the null hypothesis, if the pvalue (in Equation 2) is small. The pvalue represents the probability of observing such data table, and is calculated by the hypergeometric distribution as the following:

$$\text{pvalue} = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!}, \tag{2}$$

where notations follow those in Tabel II, and n is the total number of data. In other words, if the pvalue is small, that means the case of observing the null hypothesis is rare, therefore the feature is useful.

The reason we perform two feature ranking tests is because unlike the SVM feature ranking test, Fisher's exact test is purely data driven — the ranking only depends on the data itself, not a particular learning method.

V. Results and Discussion

We discuss hypotheses proposed in Section III-A and present the results of our analysis. We observe the consistent pattern, which supports the hypotheses that Eugenio's features are useful (H1) and play an important role (H2) to estimate the strength of agreements.

A. H1: Impact of Eugenio's features in the design problem

First, we confirm that we can estimate the strength of agreements without using keywords or contents. To compare performance difference, we use different sets of features: 1) traditional DAs only 2) traditional DAs and Eugenio's features 3) Eugenio's features only for the estimation. Table III shows the average accuracy of each algorithm. The best average accuracy in each method are 94.6% (Logistic), 90% (SVM with RBF), 66% (Kmeans) and 55% (EM with GMM). The bold texts represent the maximum average performance across folds for a given method. Note that this is the average accuracy across folds, and the standard deviation is also provided for supervised learning algorithms. The standard deviation of leave-one-out estimates is generally larger than K-fold ($K \geq 1$) cross validation, as explained in [6].

The proposed hypotheses is that Eugenio's features introduced in [7] are useful to estimate the strength of agreements. We define features are useful if any of the following is true: 1) features always improve the estimation accuracy when used with traditional DAs than traditional DAs alone (i.e. the performance is always better than without features) 2) features can replace traditional DAs (i.e. the performance is better by only using Eugenio's features) 3) features are always helpful with traditional DAs (i.e. the performance is better or the same).

All algorithms except EM with GMM achieve improvement (max 15%) in the average accuracy when Eugenio's features are used in addition or substitution to traditional DAs (i.e. last two columns in Table III). In EM with GMM, the accuracy level stays almost the same. Generally speaking, however, the Eugenio's features cannot simply replace traditional DAs. In case of SVM with RBF and Kmeans, the average accuracy degrade when only Eugenio's features are used without traditional DAs.

This analysis suggests that Eugenio's features are useful, in the sense that the estimation accuracy always improves or stays the same, when used along with traditional DAs. Note that this analysis does not tell us whether Eugenio's features are the only ones or the ultimate ones for estimating the strength of agreements. There could be another set of undiscovered features that are equally or even better help with the estimation.

B. H2: Ranking of Eugenio's features in the design problem

Next, we investigate Hypothesis 2 which states that Eugenio's features achieve higher ranking in general in feature ranking algorithms. Feature ranking algorithms rank each feature in terms of how much of information contained in each feature are useful for the estimation task. Among others, we choose SVM and Fisher's exact test to learn which features play important roles in the estimation task in both method

Rank	Feature name
1	Offer
2	**Partner decidable option**
3	**Proposal**
4	Action-directive
5	**Unconditional commitment**
6	Info-request
7	Reject
8	Open-option
9	**Unendorsed option**
10	Commit
11	Accept

TABLE V: Fisher's exact test results. The bold texts are Eugenio's features.

dependent (SVM) and method independent (Fisher's exact test) settings (details of algorithms are in Section IV-C).

Table IV shows the most useful features in SVM. The first column represents useful features for weak agreements, and the second column represents useful features for strong agreements. Three out of four, Eugenio's features (marked as bold) come up within top two most informative features in both classes, which shows the important role of Eugenio's features in the estimation task†[6].

Furthermore, we apply a method independent feature ranking algorithm, Fisher's exact test, to generalize our findings. As mentioned in Section IV-C, Fisher's exact test reports the rank of all features only based in the data itself. As shown in Table V, three out of four Eugenio's features are ranked in top five most informative features.

The consistent ranking results strongly support that Eugenio's features are not simply extra information, but they carry important information for estimating the strength of agreements.

VI. Contributions and Future work

Our approach maps the qualitative theoretical foundation for the level of joint commitment into a quantitative predictive framework. First, we confirm that Eugenio's features are useful to estimate the strength of agreements using machine learning methods. For most of the methods, the estimation performance was higher when Eugenio's features are used in addition or substitution to traditional DAs (maximum 94% accuracy). Second, we confirm that Eugenio's features play an important role in estimating the strength of agreements when compared to traditional DAs, using feature ranking algorithms. In both method dependent and method independent feature ranking algorithms, the majority of Eugenio's features come up in high ranking as the most informative features for the estimation task. Our approach can also incorporate implicit/passive acceptance of proposals by adopting Eugenio's feature, without requiring extraction of potentially sensitive information.

We would like to extend our findings regarding dialog features to other domains of problems that may exhibit different negotiation process. In particular, we would like to investigate the use of dialog features in planning problems with temporal considerations.

We envision building on this work to design an intelligent agent that observes human team planning and actively participates to plan formulation process. The agent can interject to highlight weak agreements among team members and keep track of where the team is in the planning process (i.e. what has been agreed and what has not been discussed) for more effective meetings in disaster-response planning.

[6]†Note that this feature ranking algorithm can be only done with linear kernel (not RBF) which provides in average 67% accuracy. In general, RBF achieves higher accuracy by allowing non-linear kernel as shown in Table III.

Acknowledgment

This work is sponsored by ASD (R&E) under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

Authors would like to thank Janelle Mansfield and Carrine Johnson for tagging AMI dataset.

References

[1] D. Black. On the Rationale of Group Decision-making. *Journal of Political Economy*, 56(1):23–34, 1948.

[2] K. Bousmalis, M. Mehu, and M. Pantic. Spotting agreement and disagreement: A survey of nonverbal audiovisual cues and tools. In *3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pages 1–9. IEEE, 2009.

[3] C. Cortes and V. Vapnik. Support-vector networks. In *Machine Learning*, pages 273–297, 1995.

[4] J.H. Davis. Group decision and social interaction: A theory of social decision schemes. *Psychological Review*, 80(2):97–125, 1973.

[5] D. Decoste and B. Schölkopf. Training invariant support vector machines. *Machine Learning*, 46(1):161–190, 2002.

[6] A. Elisseeff and M. Pontil. Leave-one-out Error and Stability of Learning Algorithms with Applications. In *Learning Theory and Practice*, NATO ASI Series. IOS Press, Amsterdam; Washington, DC, 2002.

[7] B. D. Eugenio, P. W. Jordan, J. H. Thomason, and J. D. Moore. The agreement process: An empirical investigation of human-human computer-mediated collaborative dialogues. *International Journal of Human Computer Studies*, 53(6):1017–1076, 1999.

[8] R.A. Fisher and S. Genetiker. *Statistical methods for research workers*, volume 14. Oliver and Boyd Edinburgh, 1970.

[9] T.S. Furey, N. Cristianini, N. Duffy, D.W. Bednarski, M. Schummer, and D. Haussler. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10):906–914, 2000.

[10] M. Galley, K. McKeown, J. Hirschberg, and E. Shriberg. Identifying agreement and disagreement in conversational speech: Use of bayesian networks to model pragmatic dependencies. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 669, 2004.

[11] S.G. Green and T.D. Taber. The effects of three social decision schemes on decision group process. *Organizational Behavior and Human Performance*, 25(1):97–106, 1980.

[12] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3):389–422, March 2002.

[13] S. Hahn, R. Ladner, and M. Ostendorf. Agreement/disagreement classification: Exploiting unlabeled data using contrast classifiers. In *NAACL*, pages 53–56, 2006.

[14] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learn.* Springer Series in Statistics. Springer New York Inc., 2001.

[15] D. Hillard and M. Ostendorf. Detection of agreement vs. disagreement in meetings: Training with unlabeled data. In *Proceedings of HLT-NAACL Conference*, 2003.

[16] S.R. Hiltz, K. Johnson, and M. Turoff. Experiments in group decision making communication process and outcome in face-to-face versus computerized conferences. *Human communication research*, 13(2):225–252, 2006.

[17] T. Jaakkola, M. Diekhans, and D. Haussler. A discriminative framework for detecting remote protein homologies. *Journal of Computational Biology*, 7(1-2):95–114, 2000.

[18] G. Ji and J. Bilmes. Dialog act tagging using graphical models. In *Proceedings of ICASSP*, volume 1, pages 33–36, 2005.

[19] C. Lottaz. Constraint solving, preference activation and solution adaptation in idiom. Technical report, 1996.

[20] C. Lottaz and I. Smith. Collaborative Design using Constraint Solving. In *Swiss Workshop on Collaborative and Distributed Systems*, 1997.

[21] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, et al. The AMI meeting corpus. In *Proceedings Methods and Techniques in Behavioral Research*, volume 88, 2005.

[22] M. Nagata and T. Morimoto. First steps towards statistical modeling of dialogue to predict the speech act type of the next utterance. *Speech Communication*, 15(3-4):193–203, 1994.

[23] E. Osuna, R. Freund, and F. Girosit. Training support vector machines: an application to face detection. In *Computer Society Conference on Computer Vision and Pattern Recognition*, pages 130–136. IEEE, 1997.

[24] C.P. Papageorgiou, M. Oren, and T. Poggio. A general framework for object detection. In *Sixth International Conference on Computer Vision*, pages 555–562. IEEE, 1998.

[25] E.L. Quarantelli. Problematical aspects of the information/communication revolution for disaster planning and research: ten non-technical issues and questions. *Disaster Prevention and Management*, 6(2):94–106, 1997.

[26] B. Scholkopf, K.K. Sung, C.J.C. Burges, F. Girosi, P. Niyogi, T. Poggio, and V. Vapnik. Comparing support vector machines with gaussian kernels to radial basis function classifiers. *Transactions on Signal Processing*, 45(11):2758–2765, 1997.

[27] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C.V. Ess-Dykema, and M. Meteer. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373, 2000.