Analysis and classification of cooperative and competitive dialogs

Uwe D. Reichel[1], *Nina Pörner*[1], *Dianne Nowack*[1], *Jennifer Cole*[2]

[1]Institute of Phonetics and Speech Processing, University of Munich
[2]Department of Linguistics, University of Illinois
{reichelu,npoerner,nowack}@phonetik.uni-muenchen.de, jscole@illinois.edu

Abstract

Cooperative and competitive game dialogs are comparatively examined with respect to temporal, basic text-based, and dialog act characteristics. The condition-specific speaker strategies are amongst others well reflected in distinct dialog act probability distributions, which are discussed in the context of the Gricean Cooperative Principle and of Relevance Theory. Based on the extracted features, we trained Bayes classifiers and support vector machines to predict the dialog condition, that yielded accuracies from 90 to 100%. Taken together the simplicity of the condition classification task and its probabilistic expressiveness for dialog acts suggests a two-stage classification of condition and dialog acts.

Index Terms: dialog acts, cooperative principle, machine learning, Gricean maxims

1. Introduction

Cooperation can be defined as "the process of groups of organisms working or acting together for their common/mutual benefit, as opposed to working in competition for selfish benefit" [1]. Grice [2] describes cooperative behavior in verbal communication in terms of the *cooperative principle* which states "Make your contribution such as it is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged" [2]. This principle can be divided into four maxims:

- Quantity: Make your contribution as informative as is required (neither less no more informative).
- Quality: Do not say what you believe to be false or for which you lack adequate evidence.
- Relevance: Contribute to the ongoing conversation topic.
- Manner: Avoid obscurity of expression and ambiguity. Be brief and orderly.

Any potential deviation from these maxims can be interpreted in two ways: first, assuming that the speaker follows the cooperative principle, the deviation gives rise to conversational implicatures [2], that is information that can be inferred from an utterance without being expressed directly nor being logically entailed. As an example, the utterance "it's cold" not embedded in a conversation on weather seems to violate the maxim of relevance unless the speaker wants to imply a request to rise temperature (close the window, turn on the heating, etc.). Second, in competitive games [3] the speaker aims to increase the own benefit at the cost of the opponent's benefit by violating the maxims and thus breaking the cooperative principle. This includes strategies as holding back information (violating the maxim of quantity), lying (violating the maxim of quality), and confounding the listener (violating all four maxims).

Related to the maxim of relevance, [4, 5] developed the *Relevance Theory*. Within this framework the *relevance* of an utterance for the hearer is defined as a function of positive cognitive effect and processing effort. The positive cognitive effect reflects the importance of the conveyed information for the hearer. The processing effort is the needed labor for the hearer to extract and make use of a conveyed information. Related to communication behavior, a cooperative speaker is expected to maximize the relevance in terms of providing important information in an easy-to-process way.

The goal of the current study is to find quantitative indications for the presence or absence of cooperative behavior and utterance relevance in cooperative as opposed to competitive dialog settings. Our approach to find evidence for the Gricean maxims is not a formal-logic [6, 7, 8] but a statistic and machine learning one. We examined several temporal and text-based parameters and used them as features for the prediction of the dialog condition by Bayesian classifiers and support vector machines.

2. Data

We used parts of the Illinois Game Corpus [9] that contains *Tangram* game dialogs by American English speakers in cooperative and competitive settings. The tangram is a puzzle consisting of seven pieces that can be combined to various shapes. Both dialog partners were separately presented with Tangram silhouettes that were reciprocally hidden from the view of the other partner. The task was to decide whether the silhouettes are the same or different by verbally describing them to each other. In the cooperative setting the partners solved this common goal in a joint effort. In the competitive setting, the partners were required to solve this task competitively, and the one solving it first was declared to be the winner. In both settings the participants were allowed to ask questions, but not to lie. Undergraduate students (ages 18-29) from the University of Illinois, all native monolingual speakers of American English, were recruited as paid participants in this study. Twelve pairs of participants took part in the experiment, some of whom were unacquainted as classmates prior to their participation. Participants were prompted to engage in free conversation for a few minutes after which they played the Tangram game together, first playing cooperatively and then competitively, with different images in each condition. Participants were seated in chairs facing one another, with no intervening table and with the printed Tangram silhouettes positioned off to the side, in front of each participant. Audio and video recordings were made on separate channels for each participant. Participants provided written consent for the use of these recordings in research.

The dialogs were manually text-transcribed and chunk-segmented, and partly manually dialog-act annotated using the

tag set of [10, 11]. This tag set was developed to describe *conversational moves*, i.e. initiations and responses with certain discourse purposes. For the given data, we augmented this label inventory by five additional tags, OFFTALK, COMMENT, COMMENT-POSITIVE, COMMENT-NEGATIVE, and UNSPECIFIED, the latter being used e.g. for laughter. The complete label set is shown in Table 4, and the description of all labels will be provided in section 5. Please consult [10] for a more comprehensive label introduction. The Dialog acts were labeled in parallel by two annotators (the second and third author), and mismatches were subsequently resolved by discussion among these two and if needed among all authors. For the current study a subset of ten dialogs by five interlocutor pairs was used, of which three were Female-Female pairs and two were Male-Female pairs. Each interlocutor pair took part in a cooperative and a competitive condition, thus our data comprises paired samples of five cooperative and competitive dialogs. Mean dialog duration amounts to 6.5 minutes.

3. Cooperativity features and prediction

In the following we introduce the different feature pools and connected hypotheses about the relation between the features and cooperative vs. competitive behavior in terms of Gricean cooperative principle and Relevance Theory. Note that the game participants were not allowed to lie so that in the given setting the maxims that may be violated are quantity, relevance, and manner. Competitive behavior is expected to be manifest primarily in the failure to provide enough and relevant information, that can be readily understood. In Relevance Theoretical terms, competitive behavior should become manifest in increasing the difficulty for the hearer to extract relevant information.

3.1. Temporal features

We extracted two basic temporal features calculated over an entire dialog: the mean chunk duration, and its standard deviation. The features are listed in Table 1. The last column indicates whether the feature showed significant differences between the cooperative and competitive condition (two-sided Wilcoxon signed rank test for paired samples). Due to the very small sample size, weakly significant results are also presented in brackets.

Hypotheses: Since cooperative behavior includes providing sufficient and relevant information (i.e. fulfilling the maxim of quantity) and adapting dialog contributions to the current needs of the dialog partner (i.e. reducing processing costs), we expect mean chunk duration and variance to be higher in cooperative than in competitive settings.

3.2. Text-based features

We extracted 9 text-based features partly inspired from the LIWC (Linguistic Inquiry and Word Count [12]) feature set and shown in Table 2: word unigram and bigram entropies, the word type-token ratio, the proportions of first and second person pronouns, the proportions of hesitations, affirmations, and negations.

Hypotheses: Again, cooperative behavior includes providing relevant and sufficient information. This higher amount of information we expect to be reflected very roughly in higher word unigram and bigram entropies, as well as in higher type-token ratios. Furthermore, we expect the proportion of personal and possessive pronouns addressing the interlocutor (*"you"*) and emphasizing the joint activity (*"us"*) to be higher in the cooperative setting, whereas the self-directed pronouns (*"me"*) should occur more often in the competitive setting. Furthermore, cooperative behavior is assumed to be reflected in higher proportions of affirmations (*"yes, ok, right,"* etc.), first since the interlocutors want to signal and not to hide successful steps to each other, and second a higher amount of successful steps should occur in cooperative settings due to a higher amount of available useful information. In contrast, competitive behavior should exhibit more negations, since due to the lack of provided information the interlocutors are forced to guess and thus make more errors reported by the dialog partner. Thus, the proportions of affirmation and negation are expected to reflect the different amounts of success in information transmission. Finally, in the competitive condition more hesitations might be observed since relevance reduction includes increased processing costs and thus a higher cognitive workload for the participants. The relation between filled pauses and cognitive workload has been observed by [13, 14].

3.3. Dialog acts

From the time-aligned dialog act annotation we derived the following three global features measured over an entire dialog: the unigram and bigram dialog act entropies, and the mutual information of subsequent dialog acts at turn transitions. Furthermore, we calculated maximum likelihood estimates for dialog act unigram probabilities separately for each dialog.

Hypotheses: Since in the cooperative setting the game participants are willing to adjust the presentation of information to the current situational needs to reduce processing costs, we expect the dialog act entropies to be higher. Furthermore, this mutual accommodation to each others' needs should be reflected in higher mutual information values of dialog act pairs at turn transitions.

3.4. Classification

Based on the dialog act probabilities and on the three feature sets introduced above we subsequently trained Bayes classifier and support vector machines for cooperativity prediction.

The Bayes classifier predicts dialog condition C (*cooperative vs. competitive*) from the dialog act sequence D by maximizing:

$$\hat{C} \quad = \quad \arg\max_{C}[P(D|C) \cdot P(C)] \tag{1}$$

The prior $P(C)$ is uniformly set to 0.5. The conditional probability $P(D|C)$ is estimated in terms of dialog act n-gram models trained on the cooperative and competitive dialogs, respectively. For this purpose we used linear interpolated bi- and unigram models and Good-Turing smoothing [15] in the form proposed in [16]. A dialog is then classified as cooperative or competitive with respect to under which of these conditions the observed dialog act sequence receives a higher probability.

Furthermore, we trained support vector machines (SVM, [17]) with a linear Kernel function on the feature pools TEMP (cf. Table 1), TEXT (cf. Table 2), and DA (cf. Table 3). The separating hyperplane was derived by sequential minimal optimization.

4. Results

Cooperativity features. For the features introduced in section 3 significant differences between the cooperative and competitive condition are indicated in the third column of Tables 1,

2, and 3 (two-sided Wilcoxon signed rank test for paired samples, $\alpha = 0.05$). Because of the small sample size 5 also weak significances at $\alpha = 0.07$ are indicated, which might turn out to become more prominent with more data. For the temporal features both hypotheses have been weakly confirmed. For the text-based features 5 out of 9 hypotheses were confirmed, and for the dialog act features none of the hypotheses.

Table 1: *Feature pool TEMP. Temporal features. The two last columns indicate the hypothesized and the observed direction of significant differences, two-sided Wilcoxon signed rank test for paired samples, $p < 0.05$, ($p < 0.07$). (Weakly) confirmed hypotheses are checkmarked.*

Feature	Description	coop vs. comp	
		hypothesis	result
dMean	mean chunk duration	$>$	$(>)$ ✓
dStd	its standard deviation	$>$	$(>)$ ✓

Table 2: *Feature pool TEXT. Text-based features. The two last columns indicate the hypothesized and the observed direction of significant differences, two-sided Wilcoxon signed rank test for paired samples, $p < 0.05$, ($p < 0.07$). (Weakly) confirmed hypotheses are checkmarked.*

Feature	Description	coop vs. comp	
		hyp.	result
hWrdUg	word unigram entropy	$>$	$(>)$ ✓
hWrdBg	word bigram entropy	$>$	$>$ ✓
ratTyptok	word type-token ratio	$>$	$(>)$ ✓
pMe	prop. "me"-type pronouns	$<$	–
pUs	prop. "us"-type pronouns	$>$	$(<)$
pYou	prop. "you"-type pronouns	$>$	$(<)$
pHes	prop. hesitations	$<$	$(<)$ ✓
pAffirm	prop. affirmations	$>$	–
pNegate	prop. negations	$<$	$(<)$ ✓

Table 3: *Feature pool DA. Dialog act features. The last column indicates the direction of significant differences, two-sided Wilcoxon signed rank test for paired samples, $p < 0.05$, ($p < 0.07$).*

Feature	Description	coop vs. comp	
		hypothesis	result
hDaUg	DA unigram entropy	$>$	$(<)$
hDaBg	DA bigram entropy	$>$	–
daMi	DA mutual information	$>$	–

Dialog act probabilities. We calculated the maximum likelihood estimates of dialog act unigrams separately for each dialog. In Table 4 the mean values over cooperative, resp. competitive dialogs are shown. This table clearly indicates that most dialog acts show a sizable distributional preference for one of the two conditions, which will be discussed in section 5.

On the global level, we measured the pairwise dissimilarity among dialogs in terms of the probabilities of the dialog acts observed in them. To calculate the dissimilarity between two dialogs their dialog act probability models P and Q are compared by their information radius $R(P||Q)$, which is given as follows:

Table 4: *Dialog act mean probabilities in cooperative and competitive settings. Distributional dialog condition preference is marked by bold face.*

Dialog act	P cooperative	P competitive
ACKNOWLEDGE	**0.1867**	0.1330
ALIGN	**0.0026**	0.0008
CHECK	**0.0552**	0.0215
CLARIFY	**0.0278**	0.0138
COMMENT	0.0407	**0.0752**
COMMENT-NEGATIVE	0.0114	**0.0264**
COMMENT-POSITIVE	**0.0512**	0.0146
EXPLAIN	**0.3234**	0.1560
INSTRUCT	0.0161	**0.0725**
OBJECT	0.0008	**0.0057**
OFFTALK	0.0099	0.0121
QUERY-WH	0.0199	**0.0460**
QUERY-YES/NO	0.0330	**0.1273**
READY	0.1282	**0.1526**
REPLY-NO	0.0059	**0.0444**
REPLY-WH	0.0198	**0.0419**
REPLY-YES	**0.0611**	0.0495
UNSPEC	0.0063	0.0066

$$R(P||Q) \quad = \quad \frac{D(P||\frac{P+Q}{2}) + D(Q||\frac{P+Q}{2})}{2} \qquad (2)$$

P and Q denote the dialog act unigram probabilities of any two dialogs from the full set of ten. The information radius distance measure fulfilling the symmetry criterion is a symmetric version of the Kullback-Leibler divergence $D(P||Q)$, which is:

$$D(P||Q) \quad = \quad \sum_{x \in X} P(x) \log_2 \frac{P(x)}{Q(x)} \qquad (3)$$

Thus $R(P||Q)$ quantifies the difference between the probabilities P and Q of the dialog acts x in dialog pairs.

As shown in Figure 1 the probability models are most similar among cooperative dialogs, and least similar among dialogs of opposite condition. The distance differences are significant for the within cooperative condition comparison opposed to the other two combinations (Kruskal-Wallis test, $\chi^2_2 = 46.3$, $p < 0.0001$, Scheffé post-hoc test, $\alpha = 0.05$) These findings indicate that speaker strategies are represented in the dialog act sequences. For dialog bigram probabilities analogous results were obtained.

Classification. We evaluated the Bayesian and support vector machine classifiers by leaving-one-out cross validation. The mean performances ranging from 90 to 100% are presented in Table 5

5. Discussion

Feature interpretation. For the temporal features, our hypotheses have been weakly confirmed. Thus in the given data, cooperation is already reflected in chunk durations. For the feature pool TEXT, however, results are less clear. The usage of pronouns and affirmations does not give any indication of cooperativity. But word token entropies and type-token ratio behaved as expected, supporting our expectation that cooperative behavior is positively correlated to the mutually submitted

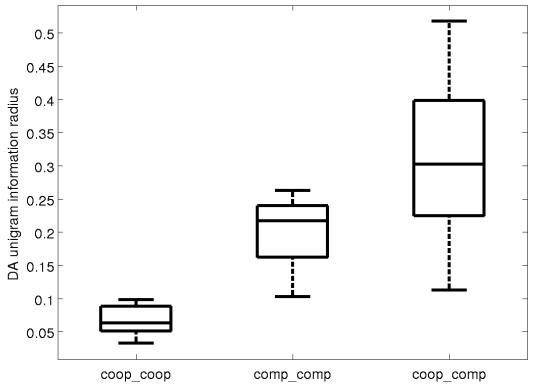

Figure 1: *Information radius between dialog act unigram probability models of dialog pairs. coop_coop, comp_comp: within condition pairings of cooperative and of competitive dialogs, respectively; coop_comp: across condition pairings.*

Table 5: *Mean classification performance by leaving-one-out cross validation in dependence on classifier and feature set. The feature sets are described in tables 1, 2, and 3.*

Classifier	Feature set	Performance (in %)
Bayes	DA n-gram model	100
SVM	TEMP	100
SVM	TEXT	100
SVM	DA	90

amount of information. Furthermore, an increased proportion of hesitations in the competitive dialogs supports the hypothesis that non-cooperative behavior increases processing costs as predicted from Relevance Theory. Also the increased proportion of negations in competitive settings as feedback to uninformed questions, is in line with our hypothesis to be a consequence of the shortage of provided information. For the feature pool DA none of the three hypothesis could be confirmed. Looking at Table 4 the lower dialog act unigram entropies are likely to be a consequence of the more unevenly distributed dialog act probabilities in cooperative settings. The dialog act most strongly related to providing information and thus to the maxim of quantity is EXPLAIN, which receives a huge amount of the total probability mass in the cooperative behavior (mean p=0.32). Since such an uneven probability distribution lowers entropy and thus works against the assumed effect of cooperative communication flexibility, it can be concluded, that for the given data dialog act entropies are not appropriate to quantify cooperativity. The same holds for the mutual information between dialog acts at turn transitions, since in cooperation EXPLAIN due to its high prior is likely to co-occur with any other dialog act, which lowers mutual information.

Selectional dialog act preferences. From the dialog act probabilities in Table 4 it can be seen that dialog acts are unevenly selected by the interlocutors across cooperative and competitive settings. In the following all quotations mark citations of the dialog act definitions given in [11].

By EXPLAIN the speaker "states information which has not directly elicited by the partner", as opposed to REPLY-*. Thus the high EXPLAIN probability in cooperative dialogs together with CLARIFY for information augmentation indicates that the speakers follow the maxim of quantity while the higher amount of REPLYs shows that the speakers do not give information deliberately and thus violate this maxim. These differences are mirrored in the differences between the CHECK and QUERY-* probabilities. Whereas CHECK refers to given or inferable information, by QUERY additional information is required. Thus the violation of the quantity maxim in competitive dialogs requires more QUERYs and less CHECKs.

COMMENTs are defined as remarks that do not add information relevant for the given task. Thus the higher amount of COMMENTs in competitive dialogs indicates violations of the maxim of relevance. The COMMENT-subclasses NEGATIVE and POSITIVE represent an online evaluation of communication situation, which is as to be expected predominantly positive for cooperative and negative for competitive dialogs.

ALIGN "checks the attention or agreement of the partner, or his readiness for the next move". The speaker thus makes sure, that the maxim of manner is not violated, and, in terms of Relevance Theory, the processing costs are low. Thus align occurs more often in cooperative dialogs.

ACKNOWLEDGE is a hearer feedback such as backchanneling, that "often ... demonstrates that the move was understood". Thus it can be used by the hearer to signal that no maxims were violated and processing cost is acceptably low. Thus as expected, its probability is higher in cooperative dialogs.

By INSTRUCT the speaker "commands the partner to carry out an action". Its higher probability in competitive dialogs might be attributed to the higher time-pressure in this condition that requires more compact command-like turns. But it can also at least partially be explained by the artifact, that the word *Tangram*, used by the speakers as an INSTRUCT move, only occurs in competitive dialogs.

Classification. Generally, dialog condition prediction turned out to be a feasible task already for basic easy-to-extract temporal and text features and for very little training data.

The high information radius values for dialog act n-gram probabilities across different dialog conditions suggest the application of dialog act probability models in Bayesian dialog condition classification which again yielded high accuracies. Conversely, this finding indicates that probabilistic approaches for dialog act tagging, as e.g. the application of Hidden Markov models [18, 19], could be improved if one trains separate n-gram models for different dialog conditions. In a two-stage classification task the appropriate dialog act probability model would then be chosen according to the dialog condition classification.

6. Conclusion

We found quantitative evidence on the temporal, text-based, and dialog act level for differences in cooperative and competitive behavior in dialogs. These differences can be theoretically anchored in terms of the Gricean cooperative principle and of Relevance Theory. For a practical application in probabilistic dialog act tagging the distinct condition-dependent probability distributions suggest the feasibility of training probability models separately for each condition, and applying the appropriate one after condition classification.

7. References

[1] Wikipedia, "Cooperation," February 11th 2015.

[2] H. Grice, "Logic and conversation," in *Speech acts*, ser. Syntax and semantics, P. Cole and J. Morgan, Eds. New York: Academic Press, 1975, vol. 3, pp. 41–58.

[3] R. Myerson, *Game Theory: Analysis of Conflict*. Harvard: Harvard University Press, 1991.

[4] D. Sperber and D. Wilson, *Relevance: Communication and Cognition*. Oxford: Blackwell, 1986.

[5] ——, "Relevance theory," in *Handbook of Pragmatics*, G. Ward and L. Horn, Eds. Oxford: Blackwell, 2004, pp. 607–632.

[6] R. Dale and E. Reiter, "Computational interpretations of the Gricean maxims in the generation of referring expressions," *Cognitive Science*, vol. 19, no. 2, pp. 233–263, 1995.

[7] N. Kadmon, *Formal Pragmatics: Semantics, Pragmatics, Presupposition, and Focus*. Oxford: Blackwell, 2001.

[8] C. Potts, *The Logic of Conventional Implicatures*, ser. Oxford Studies in Theoretical Linguistics. Oxford: Oxford University Press, 2005.

[9] "PAGE – Prosodic and Gestural Entrainment in Conversational Interaction across Diverse Languages," http:page.home.amu.edu.pl, March 22nd 2015.

[10] J. Carletta, A. Isard, S. Isard, J. Kowtko, G. Doherty-Sneddon, and A. Anderson, *HCRC Dialogue Structure Coding Manual (HCRC/TR-82)*, Human Communication Research Centre, University of Edinburgh, Edinburgh, Scotland, 1996.

[11] ——, "The reliability of a dialogue structure coding scheme," *Computational Linguistics*, vol. 23, no. 1, pp. 13–31, 1997.

[12] J. Pennebaker, M. Francis, and B. R.J., *Linguistic Inquiry and Word Count (LIWC): LIWC 2001*. New York, USA: Erlbaum, 2001.

[13] D. Barr, "Trouble in mind: Paralinguistic indices of effort and uncertainty in communication," in *Oralité et gestualité, communication multimodale, intéraction*. Paris, France: L'Harmattan, 2001, pp. 597–600.

[14] J. Arnold, C. Hudson Kam, and M. Tanenhaus, "If you say thee uh you are describing something hard: The on-line attribution of disfluency during reference comprehension," *J. Experimental Psychology: Learning, Memory, and Cognition*, vol. 33, no. 5, pp. 914–930, 2007.

[15] I. Good, "The population frequencies of species and the estimation of population parameters," *Biometrika*, vol. 40, pp. 237–264, 1953.

[16] W. Gale, "GoodTuring smoothing without tears," *J. Quantitative Linguistics*, vol. 2, pp. 217–237, 1995.

[17] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, 1995.

[18] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. Van, and E. , Meteer, "Dialogue act modeling for automatic tagging and recognition of conversational speech," *Computational Linguistics*, vol. 26, no. 3, pp. 339–373, 2000.

[19] D. Surendran and G.-L. Levow, "Dialog act tagging with support vector machines and hidden markov models," in *Proc. of Interspeech*, Pittsburgh, PA, USA, 2006, pp. 1950–1953.