Toward A Robust Dialogue System: Recognizing Dialogue Acts

Sandra Carberry
Department of Computer and Information Sciences
University of Delaware
Newark, Delaware 19716
carberry@cis.udel.edu

Abstract [1]

Dialogue acts capture the communicative action performed by a speaker, such as greeting, requesting, warning, doubting, etc. This paper explores the recognition and communication of dialogue acts. It presents research on recognizing one kind of expression of doubt in which a speaker expresses doubt at one proposition by contending that some conflicting proposition is true. It also describes research on realizing expressions of doubt effectively as natural language utterances. The paper then considers the use of machine learning to recognize dialogue acts from surface linguistic features of an utterance. However, linguistic and situational knowledge are not enough for robust recognition and communication of dialogue acts; prosodic and gestural evidence must also be taken into account. The paper argues for a multi-agent architecture that integrates linguistic, situational, gestural, and prosodic information extracted by individual evidence agents.

Introduction

It has long been recognized that utterances are not just statements that are true or false, but instead are speech acts[Aus62] by which the speaker attempts to perform an action. Allen[AP80, PA80] was the first to develop a computational model of language understanding that captured the notions of meaning[Gri69, Sea75] and action [Sea70, Sea75] espoused by Austen, Grice, and Searle. His model started from a logical representation of the speaker's utterance as a surface speech act and used plan inference techniques to derive a relationship between this surface speech act and a very limited set of possible domain goals. The path of actions connecting the speaker's utterance to this domain goal represented the speaker's partial plan for accomplishing the domain goal and served as motivation for his utterance.

Allen's seminal work inspired much further research on plan recognition and language understanding. For example, his model presumed that there were only a few high-level domain goals that the speaker might be pursuing and that the speaker's particular domain goal could be identified from a single utterance. In our early work[Car83, Car90], we noted that dialogues were generally more complex and required that the speaker's high-level domain goals and overall plan be fleshed out over several utterances. Thus we posited a *context model* that captured the speaker's domain plan as inferred from the dialogue thus far and his focus of attention in that plan. We then used plan inference rules and focusing heuristics to relate a new utterance to the system's existing beliefs about the speaker's partial plan and to expand the plan appropriately. Other researchers addressed different aspects of the plan recognition problem for language understanding[LA87, Pol86, CG93, Ram94, Les97, AZN98].

But recognition of domain goals is insufficient. Speakers also form discourse or communicative plans to exchange with their collaborative partners the information necessary to further their problem-solving goals of building a domain plan. A discourse plan contains actions that are executed in order to achieve discourse or communicative goals[Car90]; such actions are often referred to as *dialogue acts* and include seeking verification, seeking clarification, and expressing doubt.

Recognizing dialogue acts is essential if a system is to produce intelligent responses. Consider the following dialogue segment:

[1] This work was supported by NSF grants #GER-9354869 and #CDA-9703088. The research described in this paper was done with colleagues and students, particularly Keith Decker, Chandra Kambhamettu, Lynn Lambert, Ken Samuel, Leah Schroeder, and K. Vijay-Shanker.

(1) S1: *"What banks give low income mortgages?"*
(2) S2: *"State Savings gives low income mortgages."*
(3) S1: *"Doesn't State Savings require a minimum of a $50,000 down payment?"*

If utterance (3) is simply a request for verification and the queried proposition is true, then it is sufficient to affirm the proposition's truth. However, if the utterance is intended to express doubt, then an appropriate response must address the implied relationship between the queried proposition and the proposition that is being doubted.

This paper will explore the recognition and communication of dialogue acts. An expression of doubt is one kind of complex dialogue act that often initiates negotiation subdialogues and thus is very important for understanding collaborative interactions. The paper will first describe our research on a rule-based system for recognizing one kind of expression of doubt, namely utterances in which a speaker expresses doubt at a proposition P_{doubt} by contending that some other proposition P_i is true. Turning from recognition to generation, the paper will then describe our more recent work on realizing effective expressions of doubt. It will then consider the use of machine learning for recognizing dialogue acts from linguistic information contained in the words of the utterance, will discuss our research on using a randomized version of Transformation-Based Learning for dialogue act recognition, and will note the limitations of such approaches. Finally, the paper will discuss the contributions of gestural and prosodic evidence and argue for an agent-based architecture that integrates linguistic, situational, gestural, and prosodic information.

A Rule-Based Model for Recognizing Dialogue Acts

In research with Lynn Lambert[CL99], we developed a computational rule-based approach to recognizing dialogue acts in negotiation subdialogues. Such dialogues are generally initiated by expressions of doubt, and our research focused heavily on recognizing instances in which a speaker was expressing doubt at a proposition P_{doubt} by contending that some other proposition P_i was true.

Consider the following dialogue segment and the alternative possible responses by S1:

(4) S1: *"What banks give low income mortgages?"*
(5) S2: *"State Savings gives low income mortgages."*

(6a) S1: *"Doesn't State Savings require a minimum of a $50,000 down payment?*
(6b) S1: *"Does State Savings require a minimum of a $50,000 down payment?*
(6c) S1: *"Isn't State Savings located nearby?"*
(6d) S1: *"But isn't State Savings located nearby?"*

Utterance (6a) is clearly an expression of doubt at the proposition that State Savings gives low income mortgages, while (6b) seems to be merely requesting information about the size of down payments required by State Savings. So one might think that surface-negative questions of the form *Doesn't X* express doubt. However, utterance (6c) is again only a request for further information; the difference between (6a) and (6c) seems to be that banks requiring large down payments would not typically be dealing with low income customers, whereas being located nearby does not seem to have any relationship to whether a bank would provide low income mortgages. On the other hand, if we precede utterance (6c) with the cue word *But* as in (6d), then the utterance does seem to be expressing doubt. Thus recognizing expressions of doubt is not an easy task.

We developed an algorithm that could recognize when a speaker was expressing doubt at a proposition P_{doubt} by contending that some other proposition P_i is true. The algorithm incorporated knowledge from three sources:

- Linguistic knowledge: including the surface form of the utterance (such as a surface-negative question) and the presence of cue words
- Contextual knowledge: including beliefs derived from the dialogue, the structure of the dialogue, and the relative salience of different propositions in the dialogue
- World knowledge: including stereotypical beliefs that are presumed to be held by speaker and hearer and which indicate that a proposition P_i conflicts with some other proposition P_{doubt}. An example would be the stereotypical belief that banks requiring large down payments do not deal in low income mortgages.

In order to recognize that an utterance was expressing doubt at a proposition P_{doubt} due to some conflicting evidence P_i, our algorithm required evidence that:

1. the speaker has some belief in the conflicting evidence P_i

2. the conflicting evidence P_i has been brought into focus by the speaker

3. the speaker believes that the hearer believes P_{doubt} (since it is pointless to express doubt at a proposition about which there is no disagreement)

4. the speaker has not yet accepted P_{doubt}

5. the speaker believes that if P_i is true, then P_{doubt} is false

Evidence for the first belief is provided by linguistic knowledge, namely the surface form of the speaker's utterance as a surface-negative question. The next three pieces of evidence can be obtained from contextual knowledge, which indicates whether P_i was brought into focus by the current utterance, whether the other agent has claimed P_{doubt} in the preceding dialogue, and whether the speaker has already accepted P_{doubt}. The last piece of evidence is provided by world knowledge suggesting that P_i and P_{doubt} are in conflict or by the cue word *but* which generically suggests conflicting propositions. When there is equivalent evidence for interpreting an utterance as expressing doubt at one of several propositions, contextual knowledge in the form of the relative salience of the different propositions is used to arbitrate among the interpretations.

An Example

The following simple example illustrates our recognition algorithm. Suppose that we have the following two stereotypical beliefs:

(A) Banks that require large down payments do not deal in low income mortgages

(B) Banks that require applicants to have no existing debts do not deal in low income mortgages

Now consider the following dialogue segment:

(7) S1: *"What banks give low income mortgages?"*

(8) S2: *"State Savings gives low income mortgages."*

(9) S1: *"Doesn't State Savings require at least a $50,000 down payment?"*

(10) S2: *"No, State Savings no longer requires any down payment."*

(11a) S1: *"Doesn't State Savings require that applicants have no existing debts?"*

(11b) S1: *"Isn't State Savings a local bank?"*

(11c) S1: *"But isn't State Savings a local bank?"*

After utterance (10), there are two propositions that have not yet been accepted by S1 and which remain open for rejection, namely the propositions conveyed by utterances (8) and (10). Let us consider the alternative responses (11a)-(11c). The system finds all five pieces of evidence needed to recognize that (11a) is expressing doubt at the proposition that State Savings gives low income mortgages by contending that it requires the absence of existing debts. For example, the surface form of the utterance indicates that S1 has some belief that State Savings requires that applicants have no existing debts; contextual knowledge resulting from utterance (8) provides evidence that S2 believes that State Savings gives low income mortgages; and the fifth piece of required evidence is provided by the stereotypical belief that banks requiring applicants to have no existing debts do not deal in low income mortgages. On the other hand, utterance (11b) is interpreted as simply requesting verification of the location of State Savings, since the fifth piece of required evidence is missing — i.e., there is no evidence that bank location influences whether the bank requires a large down payment or provides low income mortgages. However, utterance (11c) is instead interpreted as expressing doubt, since now we have the fifth piece of evidence in the form of the cue word *but*. Since the doubt could be directed either toward the proposition in utterance (8) or the proposition in utterance (10), our system chooses the proposition in utterance (10) since it is most salient at this point in the dialogue. Thus utterance (11c) is interpreted as expressing doubt at State Savings no longer requiring a down payment by contending that it is a local bank.

Effectively Conveying Doubt

Besides recognizing doubt expressed by another agent, a collaborative system must be able to express its own doubt at claims made by the other agent, especially in situations where it has incomplete or uncertain knowledge. Analysis of naturally occurring corpora show that expressions of doubt take many different forms, from surface negative questions that query some conflicting proposition to elliptical fragments that draw attention to a particular feature of the proposition. In recent work begun with Leah Schroeder[SC00], we have been developing an algorithm for realizing an expression of doubt effectively as a natural language utterance. In this paper we will consider only instances in which the system has some belief that conflicts with the proposition posited by the other agent.

In order to respond effectively, the doubted agent must know the proposition that is being doubted since in the absence of an objection to a communicated proposition, a speaker will assume that the proposition has been accepted by the listener[CL99]. In addition, the doubted agent should also know about any conflicting evidence and the strength of the doubting agent's belief in this evidence, since such information will help the doubted agent construct a response that most effectively resolves the agents' disparate beliefs and allows the collaborative problem-solving to continue[CCC98].

So why do agents not explicitly convey all of this information? Grice's maxims[Gri75] state that a speaker should be as informative as necessary without including extraneous information. Thus in formulating an expression of doubt, we must take into account what must be explicitly communicated and what the doubted agent can be expected to infer. In addition, Clark[Cla96] has noted that speakers tend to select utterances that convey their intent efficiently, often in elliptical fragments. Since such efficiency of expression is the natural form of discourse, a listener is likely to make unintended inferences if a speaker chooses a significantly less efficient form of expression. Thus in developing an algorithm for realizing expressions of doubt, we must identify how to provide the requisite information to the doubted agent while adhering to Grice's maxims and the efficiency of expression noted by Clark.

Our approach to realizing expressions of doubt draws on the work of Vander Linden and DiEugenio[VLD96]. They developed a system for realizing negative imperatives of the form *Don't X*, *Never X*, or *Take care not to X*. They used machine learning to posit a relationship between the form of the utterance and features of the action X's relationship to the reader in terms of attention, awareness, and safety. Our work differs from theirs in several ways. Our realization algorithm must take into account the beliefs of the agents and must realize a wider variety of different forms. In addition, the examples in our current corpora do not lend themselves to machine learning in that they are often not ideal; by this we mean that in many cases the expressions of doubt did not contain all of the information that was needed by the doubted agent and thus further dialogue was needed in order to provide it. Thus in our current work, we have not used machine learning but have instead based our rules on an analysis of the dialogues and our judgements about which utterances were most successful and why.

A Realization Algorithm

Our algorithm for realizing an expression of doubt in natural language assumes that we have a model of the agents' beliefs, including stereotypical beliefs that the agents generally hold. In addition, we assume that we have a belief revision system such as that of Galliers[LRC+94] in which endorsements are used to evaluate strength of belief in a proposition. We also assume that we are given the proposition to be doubted and any conflicting evidence, though at the current time we only address instances in which there is a single piece of conflicting evidence. Our algorithm takes into account the strength of the agent's belief in P_i, the strength of belief in the implication $P_i \rightarrow \neg P_{doubt}$, and whether the conflicting evidence P_i is new information or already part of the common ground[Cla96]. The next section presents some of the rules that our algorithm uses to realize efficient and natural expressions of doubt.

Sample Realization Rules A surface negative question conveys uncertain belief in a proposition P_i. If the hearer recognizes that the speaker believes that $P_i \rightarrow \neg P_{doubt}$, then the hearer will recognize the conflict between the speaker's belief in P_i and the proposition P_{doubt}, and will thus recognize that the speaker is expressing doubt at P_{doubt} by contending that P_i is true. Therefore, a surface negative question is appropriate if the speaker believes P_i, that $P_i \rightarrow \neg P_{doubt}$, that the hearer will recognize the implication, and that P_i is more questionable than $P_i \rightarrow \neg P_{doubt}$ (since the surface negative question draws P_i into

focus and thus invites the doubted agent to address it). This leads to our first rule:

Rule-R1: **IF** the endorsement of the system's belief in P_i is at most *strong* and the strength of its belief in $P_i \rightarrow \neg P_{doubt}$ is at least as strong as its belief in P_i, **THEN** use a surface negative question that queries the truth of P_i.

Suppose that S1 has a strong belief that State Savings requires at least a \$50,000 down payment and a strong belief from default stereotypical knowledge that banks requiring large down payments do not deal with low income mortgages. Then Rule-R1 would lead to utterance (14) in the following dialogue segment (with one caveat noted below):

(12) S1: *"What bank gives low income mortgages?"*
(13) S2: *"State Savings gives low income mortgages."*
(14) S1: *"Doesn't State Savings require at least a \$50,000 down payment?"*

However, there is one problem with Rule-R1 — it doesn't say anything about the hearer recognizing that the speaker believes that $P_i \rightarrow \neg P_{doubt}$. Thus we have another rule that adds a clue word to convey the implication if the speaker does not believe that the hearer has the implication as part of his beliefs.

Rule-R2: **IF** $P_i \rightarrow \neg P_{doubt}$ is not part of the system's model of the other agent's beliefs, then initiate the expression of doubt with the cue word *but*.

Since the system has no reason to believe that the hearer believes that a bank being local would imply that it does not give low income mortgages, Rule-R2 would lead to utterance (17) in the following dialogue segment:

(15) S1: *"What bank gives low income mortgages?"*
(16) S2: *"State Savings gives low income mortgages."*
(17) S1: *"But isn't State Saving a local bank?"*

A statement of the form *"Even though X?"* conveys a relatively certain belief in X, thereby suggesting that the doubted agent will have more success in refuting the implication $P_i \rightarrow \neg P_{doubt}$ rather than the conflicting evidence P_i. This leads to the following rule:

Rule-R3: **IF** the endorsement of the system's belief in P_i is first-hand knowledge or P_i is part of the common ground of speaker and hearer, and if the system's belief in the implication $P_i \rightarrow \neg P_{doubt}$ is endorsed as strong (ie., general default beliefs), **THEN** use a statement as question of P_i preceded by the cue phrase *even though*.

Presumably, agents have first-hand knowledge of their personal information or direct experience. Thus if the system, playing the role of a financial manager, doubts that it can get a high interest rate since it has only a relatively small amount of money to invest, Rule-R3 would lead to utterance (19) in the following dialogue segment:

(18) S2: *"You can get a very high interest rate from Bank of Delaware."*
(19) S1: *"Even though I only have \$1000 to invest?"*

A simple declaration of P_i conveys a very strong belief in P_i; it also tends to convey that the speaker does not think that the hearer will be able to successfully defend P_{doubt}. Thus such a form is appropriate if the speaker has first-hand knowledge of P_i and a very strong belief that P_i implies $\neg P_{doubt}$. This leads to Rule R-4, which produces utterances that are very close to rejections:

Rule-R4: **IF** the endorsement of the system's belief in P_i is first-hand knowledge and if the system's belief in the implication $P_i \rightarrow \neg P_{doubt}$ is endorsed as very strong (ie., more than default knowledge), **THEN** use a simple declaration of P_i.

Suppose again that the system is playing the role of a financial manager and that the system has always filed only annual tax returns for its non-profit clients but files quarterly returns for its other clients. Then Rule-R4 would lead to utterance (21) below:

(20) S2: *"You need to file a quarterly tax return."*
(21) S2: *"They're a non-profit organization."*

Machine Learning for Dialogue Act Recognition

Both our recognition and realization algorithms make use of surface linguistic features such as punctuation and cue phrases. So the question arises as to

whether machine learning could be used to develop rules for recognizing dialogue acts based on surface features of the utterance. For example, an utterance such as *"That sounds good."* contains the cue words *sounds* and *good* which, together with a period at the end of the sentence and a change of speaker, suggest that the new speaker is accepting a proposal.

Machine learning has the advantage that one can give the system a large number of potentially valuable features and allow it to determine which of the features are actually useful in identifying dialogue acts. With the recent availability of corpora of dialogues, researchers have been investigating a variety of different machine learning techniques, ranging from Hidden Markov Models to decision trees[NM94b, NM94a, Mas95, RK97, PM98, WYL99, SBS+98, SCB+00], for predicting the next dialogue act or for hypothesizing the dialogue act of an utterance from features of the utterance itself and the dialogue acts of the preceding utterances.

In research with Ken Samuel and K. Vijay-Shanker, we used Transformation-Based Learning (TBL), a relatively new error-driven machine learning method devised by Eric Brill[Bri95] that achieved excellent results on part-of-speech tagging. TBL is attractive for several reasons, including the fact that its learned model consists of a set of rules that can be analyzed to derive linguistic theories.

To construct its learned model, TBL makes multiple passes through a corpus of tagged dialogues. On each pass, it examines each incorrectly tagged utterance, uses a set of predefined templates to compute all the rules that would correct the tag assigned to that utterance, and scores each rule based on how much applying that rule to the entire corpus would improve the number of correctly tagged utterances. At the end of each pass, it selects the rule with the best score, adds it to the learned model, applies the rule to the entire corpus, and begins another pass through the corpus. Learning terminates when no proposed new rule has a score exceeding some predefined threshold. The resultant learned model consists of a set of rules that are applied in sequence to assign dialogue act tags to new utterances.

Despite TBL's several attractive features, Ken Samuel (one of our project members) found that TBL was computationally intractable for dialogue act tagging. On each pass through the corpus, TBL must construct every possible rule that would correct each incorrectly tagged utterance, and the number of possible rules can grow exponentially with the number of different features included in the templates used to form the rules. Samuel devised a randomized version of TBL, called Monte Carlo TBL, that constructs only K randomly selected rules for each incorrectly tagged utterance in the training corpus[Sam98, SCVS98, CSVSW01]. Not only does Monte Carlo TBL's training time remain relatively constant as the number of features increases, but its accuracy is equivalent to standard TBL[CSVSW01].

We applied Monte Carlo TBL to the Verbmobil dialogues[REKK96, RK97] and obtained accuracy results that were statistically equivalent to the best reported results. Moreover, an analysis of the rules[Sam01] showed that TBL identified relationships between features and dialogue act tags that a human would probably fail to hypothesize. Thus we believe that machine learning offers promise as a means of correlating surface features of utterances with the dialogue act being pursued by the speaker.

However, machine learning approaches based on surface linguistic features encounter problems as the utterances become more complex and require domain knowledge such as stereotypical beliefs of the agents. For example, surface features of the utterance alone cannot differentiate surface negative questions that are expressions of doubt from those that are just requests for verification. Thus there is an upper bound on the accuracy that one can obtain from a learned model based on surface linguistic features.

Prosody and Gesture in Recognizing and Conveying Intention

Language includes all methods of communication and the instruments for communicating consist not only of the words comprising an utterance but also features of the speaker's voice, face, eyes, hands, etc.[Cla96]. Recently, researchers have begun to consider the integration of verbal and non-verbal communication[DN97, CHMN98, Nak98]. Unfortunately, the contribution of prosody and gesture to the recognition of dialogue acts has been given too little attention.

Consider the following dialogue segment:

(22) S1: *"What banks give low income mortgages?"*

(23) S2: *"State Savings gives low income mortgages."*

(24) S1: *"Isn't State Savings a small bank?"*

Human subjects vary in their interpretation of utterances that are similar to (24), and several subjects

have commented that other evidence (such as intonation or facial gesture) is needed to reliably identify the speaker's intention in such cases. The reason for this difficulty is that the speaker could believe that small banks are less likely to deal in low income mortgages, though this conflict is not as obvious as was the conflict between large minimum down payments and the provision of low income mortgages. Thus it is clear that recognizing and responding to user intention in a robust communication system should take into account gestural and prosodic evidence.

Gestural Evidence

Gesture is ubiquitous in communication. For example, humans appear to use facial gestures to communicate their attitude and intentions and appear to take the facial gestures of other agents into account in recognizing such attitudes and intentions. Although some have argued that gestures are merely facilitative (i.e., they merely help the speaker formulate an utterance), recent experiments[McN92, Cla96] indicate that they are informative.

A great deal of work has been done on incorporating appropriate gestures and postural movements into animated agents in order to make them natural and believable[PBS96, EBDP96, CNB01], since the absence of appropriate gestures can make an agent appear awkward and the inclusion of inappropriate gestures can result in confusion and miscommunication. However, relatively little work has been done on identifying how facial gesture can be used in understanding dialogue and identifying appropriate responses. Bichsel and Pentland[BP93] developed a categorization scheme for recognizing positive and negative head movement (indicating *yes* or *no*), and they noted the contribution that such gestures and facial expressions could make to a human-computer interface. Shirai[Shi96] and Iwano and colleagues[IKM+96] experimented with head movement and showed that such gestures provide information that is not present from the utterances alone. Novick and colleagues[NHRW96] investigated the role of eye gaze in a computational model of turn-taking in conversation. Quek is currently studying how gesture, speech, and gaze convey discourse structure[QAM98, Que01, QBMH01]. In [CBC96], eye blinks are detected and used as an additional cue to track faces in order to coordinate computer-mediated communication such as video conferencing.

Although much research has been devoted to recognizing emotion (happy, sad, angry, etc.) from facial expressions, facial expressions during dialogue often capture the attitude that the speaker wants to convey to the hearer, not the speaker's actual emotion[Cla96]. Moreover, work on recognizing emotion has been concerned with recognizing obvious expressions of surprise, anger, etc. and not with recognizing the more minute facial movements that characterize a speaker's attitude toward a proposition, such as the speaker's doubting a proposition. Cassell and Thorisson argue that facial expressions can convey a speaker's primary communicative intentions[CT98], and researchers have begun to consider the contribution of facial expression to recognition of user intention. For example, Hayamizu and colleagues[HHI+96] have constructed a multimodal database of spoken intentions with hand gestures and facial expressions, along with a categorization of the movements.

Although facial gestures are commonplace in human-human dialogue, one might question whether they occur in human-machine dialogue. To obtain some insight into this issue, we performed a simple experiment in which four subjects unfamiliar with our work were asked to participate in a Wizard of Oz experiment. The subjects communicated verbally with the system, and the human respondent typed her responses which appeared on the subjects' monitors. In order to provoke possible facial gestures, the human respondent occasionally provided responses that the human subjects would have reason to doubt.

Although our experiment involved too few subjects and our analysis was too informal to represent a definitive study, our experience with these few subjects provided several interesting observations. We found that the subjects did use facial gestures. In the case of expressions of doubt, we observed that the subjects often exhibited similar facial gestures, including not only the expected furrowed brows but also head tilts. These gestures generally began shortly prior to the utterance and continued into the middle of it, suggesting that not only the presence of a facial gesture but also its temporal interval of occurrence is significant in recognizing dialogue acts. The contribution of facial gesture to the recognition of dialogue acts is an area that warrants much further study.

Prosody

Early linguistic studies of prosody were concerned with identifying how different intonational features were associated with different attitudes or

beliefs. Recently, researchers have been examining how prosody might be used to aid in speech recognition and in processing discourse.

In the area of discourse processing, prosodic features have been taken into account in identifying utterance boundaries and recognizing discourse structure[GH92, HN96, NT96, Nak96]. Noguchi and Den used prosody to identify when backchannel responses might be appropriate[ND98]. Researchers have also begun to consider how prosody might contribute to understanding the communicative intention of the speaker; for example, humans typically use prosodic features to determine whether a word sequence such as *You're on Social Security* is a statement or a question. Taylor et. al.[TSI+96] developed a system utilizing statistical techniques to predict dialogue acts on the basis of prosody alone, and Wright[Wri98] used prosody along with a language model. Although it is still difficult to automatically identify prosodic events (pitch accents, phrase accents, and boundary tones), analysis-by-synthesis is a relatively new approach to the recognition of prosodic cues, which has recently shown promise for assigning JToBI labels to Japanese speech[Cam96]. In part due to the problem of automatically recognizing prosodic events, recent work with the Switchboard corpus[SBS+98, SCB+00] emphasized global features, such as utterance duration and slope of F0 contour, and showed that these provided clues for the classification of dialogue acts. However, much more research is needed on reliably extracting prosodic cues and on identifying how prosodic cues can be used in the recognition of dialogue acts with high accuracy.

A Proposed Architecture

We have presented algorithms for addressing the problems of recognizing and conveying dialogue acts, with emphasis on one particular kind of complex dialogue act — an expression of doubt. However, we have argued that reliably recognizing complex dialogue acts requires that a system take into account not only linguistic, contextual, and world knowledge but also prosodic, and gestural evidence. How should such a system be designed?

We propose a multi-agent architecture composed of independent cooperating agents that are experts at processing different kinds of evidence. A linguistic agent will be responsible for modeling dialogue structure and formulating hypotheses based on linguistic cues provided by the words in an utterance and the structure of the dialogue. A situational agent will be responsible for identifying and working with evidence from world knowledge. A gestural agent will be responsible for extracting evidence from facial gestures and head nods and constructing hypotheses based on such gestural evidence, while a prosodic agent will develop hypotheses based on intonational cues. Some of the agents will propose only a generic dialogue act (such as *express doubt*) while other agents will propose a specific instantiated dialogue act (such as *express doubt at X by contending Y*). Each agent will use whatever computational techniques are most appropriate for their particular kind of expertise. For example, the linguistic agent will most likely use machine learning while the situational agent will need to employ symbolic reasoning. However, each of these agents will associate with their hypotheses a confidence measure that captures the strength of the evidence provided by that knowledge source. An overall dialogue interpretation agent will use machine learning techniques to determine how to combine the hypotheses of the individual agents into an overall hypothesis about the intentions of the dialogue participants (such as the intention to express doubt at State Savings providing low income mortgages).

There are many reasons for maintaining separate evidence agents. It allows us to determine the impact of each kind of evidence by measuring the accuracy achieved by that agent alone and by evaluating degradation in system performance when that agent is removed. In addition, it exploits parallel processing capability and allows agents to abandon processing when other agents have provided sufficient evidence to make a decision. Moreover, additional evidence agents can easily be added in the future, such as an agent that takes into account eye gaze.

As a first step toward such an architecture, we have enhanced our linguistic agent so that it associates a confidence measure with the dialogue act tag that it assigns to an utterance[SCVS98]. Since standard Transformation-Based Learning does not associate confidence measures with its classifications, we developed a new technique for estimating the confidence of a dialogue act tag. This method draws on committee-based sampling[DE95, ED96] to produce n learned models, and the confidence measure associated with a particular classification of a new utterance is based on the number of learned models that select that classification. Our experimental results[Sam01] show that this strategy produces very good confidence measures.

Not only is a distributed agent-based architecture appropriate for recognizing dialogue acts, but such an architecture is also appropriate for realizing dialogue acts that convey a particular intention. Each of the evidence agents can suggest how they might provide some feature of the realized utterance that would contribute to conveying the desired intention, and an overall realization agent would determine which features to incorporate into the actual utterance in order to communicate the intention with high confidence that it will be correctly recognized while still adhering to the efficiency of expression noted by Clark[Cla96].

Summary

We have presented an algorithm for recognizing one kind of expression of doubt, have described an algorithm for realizing expressions of doubt effectively in natural language, and have explored our use of machine learning for recognizing dialogue acts from surface linguistic features. However, linguistic and situational evidence are insufficient for reliably recognizing dialogue acts, and systems must also take into account prosodic and gestural evidence. We argue for a multi-agent architecture composed of independent cooperating agents that are experts at processing different kinds of evidence, with an overall dialogue interpretation agent that is responsible for combining the hypotheses of the individual agents into an overall hypothesis about the intentions of the dialogue participants. Only when all sources of evidence are appropriately considered will reliable dialogue act recognition and realization be possible.

References

[AP80] James F. Allen and C. Raymond Perrault. Analyzing Intention in Utterances. *Artificial Intelligence*, 15:143–178, 1980.

[Aus62] John L. Austin. *How To Do Things With Words*. Harvard University Press, Cambridge, Massachusetts, 1962.

[AZN98] David Albrecht, Ingrid Zukerman, and Ann Nicholson. Bayesian models for keyhole plan recognition in an adventure game. *User Modeling and User-Adapted Interaction*, pages 5–47, 1998.

[BP93] M. Bichsel and A. Pentland. Automatic interpretation of human head movements. In *IJCAI Workshop on Looking at People*, 1993.

[Bri95] Eric Brill. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics*, 21(4):543–566, 1995.

[Cam96] N. Campbell. Autolabelling japanese tobi. In *Proceedings of the Fourth International Conference on Spoken Language Processing*, pages 2399–2402, 1996.

[Car83] Sandra Carberry. Tracking User Goals in an Information-Seeking Environment. In *Proceedings of The National Conference on Artificial Intelligence*, pages 59–63, Washington, D.C., August 1983.

[Car90] Sandra Carberry. *Plan Recognition in Natural Language Dialogue*. ACL-MIT Press Series on Natural Language Processing. MIT Press, Cambridge, Massachusetts, 1990.

[CBC96] Joelle Coutaz, Francois Berard, and James Crowley. Coordination of perceptual processes for computer mediated communication. In *Proceedings of International Conference on Automatic Face and Gesture Recognition*, pages 106–111, 1996.

[CCC98] Jennifer Chu-Carroll and Sandra Carberry. Collaborative response generation in planning dialogues. *Computational Linguistics*, 24(3):355–400, 1998.

[CG93] Eugene Charniak and Robert Goldman. A bayesian model of plan recognition. *Artificial Intelligence Journal*, 64:53–79, 1993.

[CHMN98] Lawrence Chen, Thomas Huang, Tsutomu Miyasato, and Ryohei Nakatsu. Multimodal human emotion/expression recognition. In *Proceedings of the International Conference on Automatic Face and Gesture Recognition*, pages 366–371, 1998.

[CL99] Sandra Carberry and Lynn Lambert. A process model for recognizing communicative acts and modeling negotiation subdialogues. *Computational Linguistics*, 25(1):1–53, 1999.

[Cla96] Herbert Clark. *Using Language*. Cambridge University Press, 1996.

[CNB01] Justine Cassell, Yukiko Nakano, and Timothy Bickmore. Non-verbal cues for discourse structure. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 106–115, 2001.

[CSVSW01] S. Carberry, K. Samuel, K. Vijay-Shanker, and A. Wilson. Randomized rule selection in transformation-based learning: A comparative study. *Journal of Natural Language Engineering*, 7(2):99–116, 2001.

[CT98] Justine Cassell and Kristinn Thorisson. Pushing the envelope: Why the communicative behaviors we notice least in animated humanoid agents matter most, 1998. submitted for publication.

[DE95] Ido Dagan and Sean P. Engelson. Committee-based sampling for training probabilistic classifiers. In *Proceedings of the International Conference on Machine Learning*, pages 150–157, July 1995.

[DN97] T. DeSilva, L.and Miyasato and R. Nakatsu. Facial emotion recognition using multimodal information. In *Proceedings of the International Conference on Information, Communications, and Signal Processing*, pages 397–401, 1997.

[EBDP96] I. Essa, S. Basu, T. Darrell, and A. Pentland. Modeling, tracking, and interative animation of faces and heads using input from video. In *Proc. of Computer Animation '96*, 1996.

[ED96] Sean P. Engelson and Ido Dagan. Minimizing manual annotation cost in supervised training from corpora. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pages 319–326, 1996.

[GH92] Barbara Grosz and Julia Hirschberg. Some intonational characteristics of discourse structure. In *Proceedings of the Second International Conference on Spoken Language Processing*, pages 429–432, 1992.

[Gri69] H. Paul Grice. Utterer's Meaning and Intentions. *Philosophical Review*, 68:147–177, 1969.

[Gri75] H. Paul Grice. Logic and Conversation. In P. Cole and J. L. Morgan, editors, *Syntax and Semantics III: Speech Acts*, pages 41–58, N.Y., 1975. Academic Press.

[HHI+96] S. Hayamizu, O. Hasegawa, K. Itou, K. Sakaue, K. Tanaka, S. Nagaya, M. Nakazawa, T. Endoh, F. Togawa, K. Sakamoto, and K. Yamamoto. Rwc multimodal database for interactions by integration of spoken language and visual information. In *Proceedings of the International Conference on Spoken Language Processing*, 1996.

[HN96] J. Hirschberg and C. Nakatani. A prosodic analysis of discourse segments in direction-giving monologues. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pages 286–293, 1996.

[IKM+96] Y. Iwano, S. Kageyama, E. Morikawa, S. Nakazato, and K. Shirai. Analysis of head movements and its role in spoken dialogue. In *Proceedings of the International Conference on Spoken Language Processing*, 1996.

[LA87] Diane Litman and James Allen. A Plan Recognition Model for Subdialogues in Conversation. *Cognitive Science*, 11:163–200, 1987.

[Les97] Neal Lesh. Adaptive goal recognition. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence*, pages 1208–1204, 1997.

[LRC+94] Brian Logan, Steven Reece, Allison Cawsey, Julia Galliers, and Karen

Sparck Jones. Belief revision and dialogue management in information retrieval. Technical report, University of Cambridge Computer Laboratory, 1994.

[Mas95] Mast, Marion and Niemann, Heinrich and Nöth, Elmar and Schukat-Talamazzini, Ernst Günter. Automatic classification of dialog acts with semantic classification trees and polygrams. In *IJCAI Workshop on New Approaches to Learning for Natural Language Processing*, pages 71–78, 1995.

[McN92] David McNeill. *Hand and Mind*. University of Chicago Press, 1992.

[Nak96] Christine Nakatani. Integrating prosodic and discourse modelling. In Y. Sagisaka, Nick Campbell, and Norio Higuchi, editors, *Computing Prosody: Computational Models for Processing Spontaneous Speech*. Springer Verlag, 1996.

[Nak98] Ryohei Nakatsu. Nonverbal information recognition and its application to communication. In *Proceedings of the International Conference on Automatic Face and Gesture Recognition*, pages 2–7, 1998.

[ND98] Hiroaki Noguchi and Yasuharu Den. Prosody-based detection of the context of backchannel responses. In *Proceedings of the International Conference on Spoken Language Processing*, 1998.

[NHRW96] David Novick, Brian Hansen, Kenneth Rubesh, and Karen Ward. Coordinating turn-taking with gaze. In *Proceedings of the International Conference on Spoken Language Processing*, 1996.

[NM94a] Masaaki Nagata and Tsuyoshi Morimoto. First steps toward statistical modeling of dialogue to predict the speech act type of the next utterance. *Speech Communication*, 15:193–203, 1994.

[NM94b] Masaaki Nagata and Tsuyoshi Morimoto. An information-theoretic model of discourse for next utterance prediction. *Transactions of Information Processing Society of Japan*, 1994.

[NT96] Shinya Nakajima and Hajime Tsukada. Prosodic features of utterances in task-oriented dialogues. In Y. Sagisaka, Nick Campbell, and Norio Higuchi, editors, *Computing Prosody: Computational Models for Processing Spontaneous Speech*. Springer Verlag, 1996.

[PA80] Raymond Perrault and James Allen. A Plan-Based Analysis of Indirect Speech Acts. *American Journal of Computational Linguistics*, 6(3-4):167–182, 1980.

[PBS96] C. Pelachaud, N. Badler, and M. Steedman. Generating facial expressions for speech. *Cognitive Science*, 20(1):1–46, 1996.

[PM98] Massimo Poesio and Andrei Mikheev. The predictive power of game structure in dialogue act recognition: Experimental results using maximum entropy estimation. In *Proceedings of the International Conference on Spoken Language Processing*, 1998.

[Pol86] Martha Pollack. A Model of Plan Inference that Distinguishes Between the Beliefs of Actors and Observers. In *Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics*, pages 207–214, New York, New York, 1986.

[QAM98] Francis Quek, Rashid Ansari, and David McNeill. Gesture, speech, and gaze in discourse management, 1998. Project Description.

[QBMH01] Francis Quek, Robert Bryll, David McNeill, and Mary Harper. Gestural origo and loci-transitions in natural discourse segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2001.

[Que01] Francis Quek. Instrumental access to natural multimodal discourse. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, 2001.

[Ram94] Lance Ramshaw. Correcting real-word spelling errors using a model of the problem-solving context. *Computational Intelligence*, pages 185–211, 1994.

[REKK96] Norbert Reithinger, Ralf Engel, Michael Kipp, and Martin Klesen. Predicting dialogue acts for a speech-to-speech translation system. Technical Report Verbmobil-Report 151, DFKI GmbH Saarbrücken, 1996.

[RK97] Norbert Reithinger and Martin Klesen. Dialogue act classification using language models. In *Proceedings of EuroSpeech-97*, pages 2235–2238, 1997.

[Sam98] Ken Samuel. Lazy transformation-based learning. In *Proceedings of the Eleventh International Florida Artificial Intelligence Research Symposium Conference*, pages 235–239, 1998.

[Sam01] Ken Samuel. *Discourse Learning: An Investigation of Dialogue Act Tagging Using Transformation-Based Learning.* PhD thesis, University of Delaware, January 2001.

[SBS+98] Elizabeth Shriberg, Rebecca Bates, Andreas Stolcke, Paul Taylor, Daniel Jurafsky, Klaus Ries, Noah Coccaro, Rachel Martin, Marie Meteer, and Carol Van Ess-Dykema. Can prosody aid the automatic classification of dialog acts in conversational speech? *Language and Speech: Special Issue on Prosody and Conversation*, 1998.

[SC00] Leah Schroeder and Sandra Carberry. Realizing expressions of doubt in collaborative dialogue. In *Proceedings of the 18th International Conference on Computational Linguistics*, pages 740–746, 2000.

[SCB+00] Andreas Stolcke, Noah Coccaro, Rebecca Bates, Paul Taylor, Carol Van Ess-Dykema, Klaus Ries, Elizabeth Shriberg, Daniel Jurafsky, Rachel Martin, and Marie Meteer. Dialogue act modelling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–373, 2000.

[SCVS98] Ken Samuel, Sandra Carberry, and K. Vijay-Shanker. An investigation of transformation-based learning in discourse. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 497–505, 1998.

[Sea70] John R. Searle. *Speech Acts: An Essay in the Philosophy of Language.* Cambridge University Press, London, England, 1970.

[Sea75] John R. Searle. Indirect Speech Acts. In Peter Cole and Jerry Morgan, editors, *Syntax and Semantics: Speech Acts*, volume 3, pages 59–82. Academic Press, Inc., New York, New York, 1975.

[Shi96] Katsuhiko Shirai. Modeling of spoken dialogue with and without visual information. In *International Symposium on Spoken Dialogue*, pages 101–104, 1996.

[TSI+96] P. Taylor, H. Shimodaira, S. Isard, S. King, and J. Kowto. Using prosodic information to constrain language models for spoken dialogue. In *Proc. of the International Symposium on Spoken Dialogue*, pages 129–132, 1996.

[VLD96] Keith Vander Linden and Barbara DiEugenio. A corpus study of negative imperatives in natural language instructions. In *Proceedings of the 16th International Conference on Computational Linguistics*, 1996.

[Wri98] Helen Wright. Automatic utterance type detection using suprasegmental features. In *Proceedings of the International Conference on Spoken Language Processing*, 1998.

[WYL99] Chung-Hsien Wu, Gwo-Lang Yan, and Chien-Liang Lin. Speech act modelling in a spoken dialogue system using fuzzy hidden markov model and bayes decision criteria. In *Proceedings of Eurospeech'99*, 1999.