Causes and Strategies for Requesting Clarification in Dialogue

David Schlangen
Universität Potsdam
das@ling.uni-potsdam.de

Abstract

We do two things in this paper. First, we present a model of possible causes for requesting clarifications in dialogue, i.e., we classify types of non-understandings that lead to clarifications. For this we make more precise the models of communication of (Clark, 1996) and (Allwood, 1995), relating them to an independently motivated theory of discourse semantics, SDRT (Asher and Lascarides, 2003). As we show, the lack of such a model is a problem for extant analyses of clarification moves. Second, we combine this model with an extended notion of "confidence score" that combines speech recognition confidence with different kinds of semantic and pragmatic confidence, and argue that the resulting processing model can produce a more natural clarification and confirmation behaviour than that of current dialogue systems. We close with a description of an experimental implementation of the model.

1 Introduction

It is widely accepted that it would be desirable for dialogue systems to be able to produce and understand the whole range of *Clarification Requests* (CRs) that can be found in human-human dialogue, as exemplified in the following:

(1) a. A: I talked to Mary-Ann Parker-Tomlison.
B: Parker-WHO?

b. A: Well, I've seen him.
B: Sorry, you *have* or you *haven't*?

c. A: Did you talk to Peter?
B: Peter Miller?

d. A: Did you bring a 3-5 torx?
B: What's that?

A precondition for fulfilling this desideratum is a detailed analysis of the communication problems that lead to the need for clarification. As we show in this paper, extant approaches to CR do not satisfy this precondition. We propose that a good starting-point for developing a more general analysis is a multi-levelled model of communication along the lines of (Clark, 1996) and (Allwood, 1995), distinguishing (among other things) between *acoustic understanding* and *semantic understanding*.[1] We explore such a model from the perspective of generating and interpreting CRs, making the central concepts of the model precise by relating it to an independently motivated model of discourse semantics called SDRT (Asher and Lascarides, 2003).

Deciding on whether to produce CRs is part of the *Confirmation Strategy* (CS) of a dialogue system (cf. (San-Segundo et al., 2001), *inter alia*). An *explicit confirmation* of an understanding can be sought via a CR, whereas *implicit confirmation* can be sought by displaying the system's understanding:

(2) a. *Explicit confirmation*: "Did you say you want to leave from Potsdam?"

b. *Implicit confirmation*: "From Potsdam. To where?"

Current dialogue systems base their decision on the CS to follow only on their confidence in the speech recognition results. It would be desirable, however, if they could clarify or confirm other hypotheses as well, for example about reference resolution, depending on their confidence

[1] Several recent papers (Gabsdil, 2003; Larsson, 2003) have followed a similar approach, but with a somewhat narrower focus. (Gabsdil, 2003) is mostly concerned with CRs reacting to speech recognition, while (Larsson, 2003) offers a similar, but less fine-grained classification and deals more with integrating CRs into a specific kind of dialogue management strategy.

in that resolution:

(3) User: Send the file to Peter.
System_a: Do you mean Peter Miller?
System_b: Will send the file to Peter Miller. Anything else?

Moreover, confidences on different levels of processing should be allowed to interact. In a situation where the speech recogniser cannot decide between the hypotheses "Sandy" and "Andy" for a certain input, but where the former proper name can be resolved to a more salient discourse referent than the latter, a dialogue system should ideally prefer the former hypothesis and choose implicit confirmation (variant A below), rather than explicitly clarifying which alternative to choose (variant B):

(4) User: Send the file to {Sandy | Andy}.
Sys_a: To Sandy, OK.
Sys_b: Did you say *Sandy* or *Andy*?

The remainder of the paper is organised as follows. After presenting an initial classification of CRs and discussing extant approaches in the next section, we propose in Section 3 a model of causes for requesting clarification. Building on this theoretical model we turn in Section 4 to extending the concept of *confidence* in a hypothesis in order to produce the CS behaviour sketched above. We also discuss initial findings from an experimental implementation of this idea.

2 Clarification Requests

2.1 A First Classification

The examples in (1) above together with (5) below illustrate the wide range of CRs that can occur in dialogues, varying with respect to their *form* (from conventional forms (5-a) to full sentences ((1-d),(5-c-i) and (5-c-ii)), to sentential fragments ((1-a),(1-b),(1-c),(5-b-i) and (5-b-ii))) and with respect to their *function* (clarifying acoustic understanding ((5-a) and possibly those in (1)); reference ((1-c),(1-d),(5-b-ii) and possibly (1-a) and (5-b-i)); or pragmatic impact (the examples in (5-c))).

(5) a. A: Did you talk to Peter?
B: Sorry? / Pardon? / You what?
b. (i) A: Did you bring a 3-5 torx?
B: A what?
(ii) A: George Bush is in hospital.
B: Junior or senior?
[from (Gabsdil, 2003)]
c. (i) A: What time is it?
B: Do you want to leave?
(ii) A: Can you pass me the salt?
B: Is that a question or a request?

What the questions in these examples have in common is that, unlike "normal" questions, they are not about the state of the world in general, but rather about aspects of previous *utterances*: they indicate a problem with *understanding* that utterance, and they request repair of that problem. We take this to be the defining features of CRs. Note that this definition includes correctional uses of CRs as illustrated in (6); in this case the problem is taken to originate on the side of the speaker of the original utterance rather than on the side of the CR initiator.

(6) a. A: Dear police men....
B: Police *men*?
A: Alright then, police *people*.
b. Student: 3 + 4 = 8
Teacher: 3 + 4 = *8*?

We will focus on the possible *functions* of CRs in this paper, leaving the question of how to map CR *form* to that function to further work.[2] We simply observe at this point that some CRs indicate the kind of understanding problem that occurred; e.g., in (1-c) and (5-b-ii) this seems to be a problem with identifying the intended referent; in (1-d) and (5-b-i) a lexical problem; in (5-c) a problem with recognising the intention behind the utterance. This observation will form the basis of the classification developed below in Section 3, where we further develop extant models to make the pre-theoretic notion of *understanding* precise.

Contrasting (1-c) and (5-b-ii) with (7) below illustrates another dimension for classification. Where the former two CRs ask for a *confirmation* of a hypothesis, the latter asks for a *repetition* (or *reformulation*) of the problematic element.

(7) A: I talked to Shanti.
B: WHO?

We call this dimension *severity*; this represents the intuition that a problem that leads to a request for repetition is more severe than one that leads to a request for confirmation; we will make this notion of *severity* precise in Section 4. Note that a *confirmation request* can be realised as an alternative question, as in (5-b-ii) and (4)_b, or as a y/n-question, as in (1-c).

Lastly, we also distinguish between CRs that point out a problematic element in the original utterance, and those that don't. The former category is illustrated by the CRs in (1) and (5-b), the latter by those in (5-a) and (5-c). We call this dimension *extent*.[3]

[2] (Purver et al., 2001) investigates such a mapping, based on the classification discussed below in Section 2.2.

[3] The dimensions for classification introduced here are related to, but different in some aspects from those used in (Larsson, 2003). Our term 'CR' covers what Larsson calls negative feedback as well as what he calls checking feedback, whereas

Before we finally come to the description of the dimension *level of understanding* in the next Section, we will briefly look at an earlier analysis of CR that does not make these distinctions.

2.2 Previous Analyses

In a number of papers (Ginzburg and Cooper, 2001; Purver et al., 2001), Jonathan Ginzburg and colleagues have developed an influential analysis of CR. The authors define two readings that can be ascribed to CRs, which they name the *constituent* reading and the *clausal* reading.[4] (8-b) shows paraphrases of these readings for the CR in (8-a).[5]

(8) a. A: Did Bo leave? — B: Bo?
b. *clausal*: Are you asking whether Bo left?
constituent: Who's Bo?

These readings are defined informally by (Ginzburg and Cooper, 2001) (henceforth G&C) as follows: the *clausal reading* "takes as the basis for its content the content of the conversational move [$\approx$ speech act, D.S.] made by the utterance being clarified. [It] corresponds roughly to 'Are you asking / asserting that X?', or 'For which X are you asking / asserting that X?'."; the *constituent reading* is a reading "whereby the content of a constituent of the previous utterance is being clarified."

Let's look at the conditions under which a dialogue participant (be that a dialogue system or a human) might intend one or the other reading. We begin with the situation shown in (9-a), taken from G&C.

(9) a. A: Did Bo leave? — B: Who?
b. *clausal*: For which x are you asking whether x left?
constituent: Who's Bo?

It seems that the clausal reading is appropriate both in situations where A failed to recognise the name acoustically as well as when she failed to resolve the reference. An answer to this question will always resolve both kinds of problems; i.e., this reading does not make a difference between these kinds of understanding problems. The constituent reading, on the other hand, does, and is only appropriate for repetition requests targeting the semantic/pragmatic problem 'reference resolution.'

The next example, also from G&C, shows a CR that has the form of a reformulation of the original content. In this case a constituent reading is not available in G&C's analysis, since they postulate a phonological-identity condition for constituent readings which is violated here.

(10) a. A: Did Bo leave? — B: My cousin?
b. *clausal*: Are you asking whether my cousin left?
constituent (not allowed by G&C): Is the denotation of "my cousin" the denotation of "Bo"?

However, in this example the intended distinction between the readings seems to break down, since the function of the clausal reading here is to clarify a hypothesis about the denotation of a referring expression, which in other cases is the function of constituent readings.

To summarise, G&C's analysis does not explicitly record the problem that leads to the need for clarification. This leads to a loss of information; information which however will be present on the side of the CR producer (who knows where the understanding problem occurred) and presumably should be present on the side of the recipient, who might want to react differently depending on the assumed problem. It seems that the construct "Are you X-ing whether p?" is too general to make these fine-grained distinctions.[6]

severity gives a finer classification of what he calls 'eliciting feedback'. Larsson gives a classification comparable to our *extent* dimension only indirectly, via a classification of the *forms* used to express CR as syntactically fragmental or complete; however, since syntactically complete utterances can nevertheless target individual elements ("Which Peter are you talking about?"), these categorisations are not congruent.

[4] They also mention that a third reading might be needed, namely a *lexical* reading in which "the *surface form* of the utterance is being clarified, rather than the content of the conversational move. This reading therefore takes the form 'Did you utter X?'." (Purver et al., 2001). However, the authors do not offer a formalisation of this reading, so we will concentrate on the two readings for which they do.

[5] These readings are realised technically by a straightforward formalisation of these paraphrases in an HPSG framework, using an *illocutionary-act* relation for the clausal reading and a relation *content* for the clausal readings, where both relations take signs as arguments. Since the formalisation is so close to the paraphrases (and is in any case not backed up by a formal semantics of the predicates used), we can use in the following arguments just the paraphrases without missing crucial details.

[6] There is also a technical problem with this analysis, which can be illustrated with the following example.

(i) A: Can you pass me the salt? — B: The salt?

A's utterance is of course an example of an indirect speech act. Since G&C assume that the illocutionary force of the previous utterance is represented in the CR-reading ("Are you asking/asserting/etc.-ing whether...") generated by the grammar, they have to find a way to capture this indirectness. There are only two, equally unattractive, options to do this: either the authors have to assume that the grammar directly assigns A's utterance the force *request* (rather than *question*), so that the clausal reading of B's utterance can be paraphrased as "Did you *request* that I pass the salt?". But interpreting indirect speech acts is a highly context-dependent task and not something that can be decided on syntactical grounds alone. The other option is to stick with the speech act type that is normally associated with interrogatives, and arrive at a reading that can be paraphrased as "Are you asking whether I can pass you the SALT?". This, however, is presumably in most cases not the right interpretation of

Level	Clark	Allwood	Ginzburg *et al.*
4	proposal & consideration	reaction to main evocative function	
3	meaning & understanding	understanding	clausal reading; constituent reading
2	presentation & identification	perception	lexical reading
1	execution & attention	contact	

Figure 1: The four basic levels of communication

We now turn to exploring a model that does make these distinctions.

3 A Model of Causes for Clarification

3.1 The Fundamental Distinctions

Herb Clark (Clark, 1996) and Jens Allwood (Allwood, 1995) independently developed a model of the (hierarchically ordered) tasks involved in communication, as shown schematically in Figure 1.[7] (We have also assigned the readings defined by G&C to the appropriate levels in the last column.) This model can serve as a basis for classifying the function of CRs. For example, the CRs shown in (11) can be classified as each targeting a different level according to the model.

(11) a. [You are sitting in a subway train when A sits down on the seat next to you, talking. You might say:] Are you talking to me?
b. A: I saw Peter.
B: What did you say? (I didn't hear you.)
c. A: I saw Peter.
B: Which Peter?
d. A: My mother is a lawyer.
B: Are you trying to threaten me?

The distinctions made by this model, however, are still fairly coarse-grained. It seems desirable to further analyse the levels—and especially the third one, that of 'meaning and understanding"—so as to capture for example the difference between the CRs in (12) below. We will do this in the next section.

(12) a. A: I ate a Pizza with chopsticks the other day.
B: A Pizza with chopsticks on it?
b. A: Please give me the double torx.
B: What's a torx?
c. A: Please give me the double torx.
B: Which one?
d. A: Every wire has to be connected to a power source.
B: Each to a different one, or can it be the same for every wire?

3.2 A More Fine-Grained Model

How shall we further carve up the level "meaning & understanding"? One well-known additional distinction that seems useful is that between *literal meaning* and *speaker meaning*.[8] For instance, this distinction is evoked in the following categorisation given by (Larsson, 2003):

- *Semantic Meaning:* discourse-independent meaning. E.g. word meanings.
- *Pragmatic Meaning:* domain-dependent and discourse-dependent meaning, further split into:
  - referential meaning, e.g. referents of pronouns, temporal expressions;
  - pragmatic meaning proper: the relevance of u in the current context.

This points in the right direction, but still needs to be made more precise: for instance, what is "relevance" here? Where do the examples in (12) fit in that schema?

To make these terms precise, and to add further distinctions, we have devised a model that is closely inspired by how the discourse semantics theory SDRT (Asher and Lascarides, 2003) sets up the syntax/semantics/pragmatics interface. In particular, we use the idea of using *semantic underspecification* to allow for "pragmatic intrusion" into the determination of the truth value of an utterance; where roughly the underspecified logical form generated by the grammar corresponds to the (set of) literal meaning(s) of an utterance, and the pragmatically resolved LF to the speaker meaning. We also use SDRT's idea of spelling out contextual relevance as the need for determining a rhetorical relation with which to connect a new utterance to the context, and the concept of speech act related goals (SARGs), i.e. goals that are conventionally connected to certain types of speech acts.[9] (These ideas will be illustrated with examples below in Section 4 when we discuss the implementation.) This move allows us to say precisely what constitutes an understanding problem in this model—namely, a failure to tackle one (or more) of the precisely defined tasks.

the CR.

[7] There are significant differences between these two models; however, for our purposes the similarities dominate and so we treat the models as terminological variants; keeping in mind that we simplify both models considerably.

[8] Which goes back at least to Grice's studies from the 1960s reprinted in (Grice, 1989).

[9] In this paper we concentrate on outlining how this theory could be used to model clarification and leave further formalisation for future work.

Level	Description
1	establishing contact
2	speech recognition
3a	parsing:
3aa	recognising all words
3ab	determining syntactic structure
3ac	determining a *unique* syntactic structure
3b	resolving underspecification:
3ba	reference
3bb	tense, scope, presuppositions, lexical ambiguities, etc.
3c	contextual relevance, computing the rhetorical connection
4	recognising speaker's intentions; evaluating resulting discourse structure

Figure 2: The fine-grained model

The resulting fine-grained model is shown in Figure 2; the additional levels are further motivated by the examples of CRs shown in (13), which indicate problems on each of these levels (the labels refer to the labels of the (sub-)levels in the figure).[10,11]

(13) a. *3aa*: see e.g. (1-d) or (5-b-i) above, or:
A: Peter kowtowed again.
B: What does "kowtow" mean?

b. *3ab*:
A: The cat that the mouse that the flee bit saw slept.
B: Who did what? Again, please.

c. *3ac*:
A: I saw a man with a telescope.
B: What do you mean? Did you see a man who was holding a telescope, or did you use a telescope to watch him?

d. *3ba*: see e.g. (1-c), (12-c) above.

e. *3bb*: see (12-d) above, or:
A: I went to the bank yesterday.
B: As in "monetary institute", or in "going fishing"?

f. *3c*:
[At court, A being a witness and B the judge.]
A: Max fell. John pushed him.
B: Witness, do you mean that he fell *because* he was pushed by the defendant?

g. *4*:
(i) A: My mother is a lawyer.
B: Are you trying to threaten me?
B′: Why are you telling me that?
B″: What do you want to say with that?

(ii) A: Let's meet next week.
B: My parents in law are visiting on Tuesday.
A: So are you saying that Tuesday is good or bad?

To summarise what we have done so far, we have shown a model that can distinguish in a fine-grained manner possible problems during the processing ("understanding") of an utterance which can lead to the need for clarification.[12] This models one of the dimensions for classification which we described in Section 2. What remains to be done is to explain how the problems at each level can be of different *severity*, leading to either *repetition* or *confirmation requests*. This we will do in the next section, where also a general strategy for dealing with processing hypotheses will be discussed.

4 Clarification Strategies

4.1 Extending the Concept of Confidence Scores

In current spoken dialogue systems there is a very dominant source for understanding problems: speech recognition (SR). Many existing dialogue systems (cf. e.g. (San-Segundo et al., 2001)) make use of the confidence scores returned by SR systems together with each recognition hypothesis. Based on this value the system can decide whether to reject the hypothesis (which will lead to a repetition request, e.g. "Can you repeat?", "Pardon?"), whether to confirm it (explicitly with a confirmation request or implicitly), or whether to accept it without generating explicit feedback. This strategy is represented schematically in Figure 3, where the different CSs are distributed over the space of possible confidence values.

The idea we want to explore here is very straightforward: this concept of *confidence score* should be extended to all levels of processing. At all processing stages where the system has to rely on "guesses" (nonmonotonic inferences, heuristics, etc.), it should assign a confidence value to its hypothesis. These confidence

[10] Note that despite the rather technical names for the levels this is still a theoretical model of the process of understanding—one, however, that could be implemented in a dialogue system (see (Schlangen et al., 2001) or Section 4 below). Note also that we do not want to make any claims about the psychological status of these postulated levels. All we claim here is that these levels (which are independently motivated in SDRT) are useful for distinguishing types of CRs.

[11] These CRs are not necessarily very natural sounding; the point is just that one can construct CRs that target exactly those postulated levels. The examples are presented here to theoretically motivate those levels. We are currently conducting a detailed corpus analysis to determine the coverage that can be achieved with this model.

[12] This model is backed up by an independently-motivated formal semantic theory, which however for reasons of space we cannot present here in any detail; cf. (Asher and Lascarides, 2003) for this.

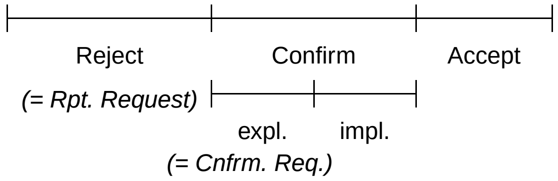

Figure 3: Confirmation Strategies

values should then be combined in some principled way (for example by taking a weighted average), to determine the Confirmation Strategy (repetition or confirmation request, implicit confirmation, acceptance) and the level of processing that is to be indicated as the primary source of the problem.[13]

In the simplest case, a system should be able to delay its decision on a CS until processing of (a certain number of) the hypotheses is completed. For example, imagine the case of a travel information system based in the UK. This system might be able to offer *flights* to Boston (and so should be able to recognise the input "Boston"), but it also knows that there are no *trains* from the UK to Boston. So in this case it should be able to make a decision for one of the hypotheses in the following example.

(14) Sys: Hello. This is TravelUK. How may I help?
User: I want to go to {Brighton | Boston} by train.

The strategy to make use of combined confidence scores could be implemented in many different systems; for example even the rather simple technique of reference resolution via salience lists has an inherent quantitative element that could be used. However, to make the idea more precise, we will in the next section describe an experimental implementation of it in a dialogue system that follows the approach to discourse interpretation described in the previous section. Such a system should ultimately be able to produce the whole range of CRs according to the dimension *level of processing* as discussed in Section 3.2 as well as along the dimension *severity*.

4.2 An Experimental Implementation

So far we have said very little about how the theory of discourse semantics alluded to in Section 3.2 tackles its various tasks. In a nutshell, the theory SDRT can be seen as a combination of *dynamic semantics* (e.g., DRT, (Kamp and Reyle, 1993)) plus (AI-based) pragmatics. In contrast to traditional dynamic semantics, SDRT attempts to represent the *pragmatically preferred* interpretation of a discourse. The central notion of *Discourse Update* is formulated in SDRT within a precise nonmonotonic logic, in which one computes the rhetorical relation (or equivalently, the speech act type) which connects the new information to some antecedent utterance. This speech act places constraints on content and the speech act related goals or SARGs; these in turn serve to resolve semantic underspecification. Note that those SARGs are goals that are either conventionally associated with a particular type of utterance or are recoverable by the interpreter from the discourse context; this distinguishes the goals that interact with linguistic knowledge from goals in general.

The implementation of the theory which we extended for this paper, RUDI (Schlangen et al., 2001; Schlangen, 2003), works in the domain of appointment scheduling (we will refer to the extended version as RUDI$_{clar}$). It focuses on resolving one particular kind of underspecification, namely that arising from the need to "bridge" definites to their context. To give an example, for (15) the system computes that the "Wednesday afternoon" is 'bridged' via the relation 'next' to the time of utterance:

(15) A: What is a good time for you in the next couple of weeks?
B: Wednesday afternoon would be good.

It does this by non-monotonically inferring the rhetorical relation connecting the second to the first utterance (*Question-Answer Pair*), and using constraints on this relation (roughly, times mentioned in the answer must overlap with that from the question) to resolve underspecification. Before we further describe how the algorithm works, however, we give a couple of examples illustrating the clarification behaviour of the system.[14]

Example (16) shows an input where the SR component offers two hypotheses that both have to be considered.[15] Let's assume that there is no salient Monday the 13th given the dialogue context. In such a situation we want the system to dramatically lower its confidence in the Monday hypothesis, leading to a situation where only the Sunday hypothesis will have to be confirmed, and only *implicitly*, rather than having to clarify both hypotheses.

[13] This is a generalisation of the approach taken for example by (Walker et al., 2000), who use the output of the semantic and pragmatic modules of their dialogue system to dramatically improve the classifier that judges whether a SR hypothesis is correct or not, compared to a classifier that just uses SR-features.

[14] We should stress that RUDI$_{clar}$ is a proof-of-concept implementation of the model presented here and not a proper dialogue system. Neither is the system actually connected to a speech recogniser—we simulate this by annotating the input with confidence scores—nor does it generate the clarification *forms* shown in the examples—rather it produces abstract instructions of the form "confirm element x", "request repetition of element y" etc. which could be used in such a generation. Moreover, RUDI models an overhearer of a conversation, not an actual participant. In RUDI$_{clar}$ this is an overhearer that barges in if it feels the need to clarify something.

[15] At present the system can only handle two alternative hypotheses, where only one constituent may differ.

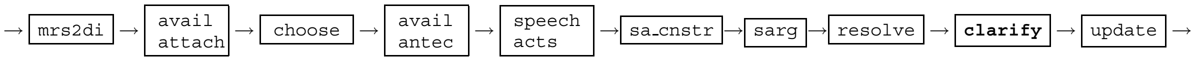

Figure 4: Schematic Overview of RUDI$_{clar}$

(16) A: I'm free on $\{$Sunday $\langle 55 \rangle|$ Monday $\langle 45 \rangle\}$ the 13th.

Example (17) brings in combined confidence values in a more subtle way.

(17) A: Let's meet this weekend.
B: How about $\{$Sunday $\langle 57 \rangle|$ Monday $\langle 43 \rangle\}$?

Here it is not the case that one hypothesis is ruled out completely; the difference between the hypotheses here is that one is 'costlier' to maintain. In our approach the dialogue "Let's meet next weekend. – How about Monday?" is just about coherent, under a reading where the second utterance indirectly corrects the plan (a more explicit version of this would be "How about Monday *instead*?"). This speech act (*Plan-Correction*), however, is inferred in RUDI$_{clar}$ with a much lower confidence than the more direct speech act *Plan-Elaboration* that is computed for the Sunday-variant of B's utterance. Hence, the system prefers that latter variant and proceeds accordingly.

The previous examples combined an ambiguity introduced by the SR module with confidence scores from further levels. However, new ambiguity can also arise during these latter processing stages. In (18) the temporal expression in B's utterance cannot be uniquely resolved (and in this sense the utterance is actually slightly incoherent, since it violates the uniqueness presupposition of definites), and so the system has to clarify the intended reference.

(18) A: Let's meet this weekend.
B: How about at 3pm?
RUDI: 3pm on Saturday or 3pm on Sunday?

The last example, (19), shows another source for quantified hypotheses: resolving any temporal expression other than "tomorrow" to the next day is dispreferred in the system, and so its confidence in this resolution is lowered and it has to be clarified.

(19) A [on a Friday]: Let's meet this weekend.
B: How about Saturday?
RUDI: You mean tomorrow?

We now sketch how the system works. Reflecting the modularity of the underlying theory, RUDI($_{clar}$) divides the update process into several stages (shown schematically in Figure 4). The initial module `mrs2di` postprocesses the semantic representation provided by the grammar, for example by including underspecified bridging relations for definites. RUDI$_{clar}$ allows logical forms to be annotated with confidence values (following an approach similar to that of (Gabsdil and Bos, 2003), associating the confidence values with labels in an underspecified LF), and it allows alternative hypotheses as input (only two at present). In this way we can represent in the system a situation where the speech recogniser cannot make a decision, as in (16) or (17) above.

At the next stage, an utterance in the context is chosen to which the current one can be attached via a rhetorical relation, and this in turn determines which antecedents for bridging are available. (Should this choice turn out to lead to failure in successive modules, the system can backtrack and choose another attachment site.) The speech act(s) of the current utterance is (are) then inferred non-monotonically (if there is more than one hypothesis coming from the previous step, this is done for each of them) from information about the antecedent and the current utterance and axioms for each relation. The next module, `sa_cnstr`, tests whether certain constraints on the meaning of the speech acts are satisfied by the utterances that are being connected. After this, the SARGs are computed and any remaining underspecification is resolved.

Finally, a new module *clarify* compares (if there is more than one) and scores the hypotheses, assigning scores for the bridging decisions and for the speech-act inferences. Some of the rules used here are shown in Figure 5. A weighted average is computed, and, based on the resulting score, the module decides on whether to launch into a clarification sub-dialogue, and if so, which clarification strategy to follow. (We set the thresholds for this and the weights for the average manually to achieve the behaviour described here; see discussion below.) The level at which RUDI$_{clar}$ targets the clarification is always the lowest one where there was a problem; i.e., where alternative hypotheses were introduced, or where no result could be computed. For instance, in (17) this would be the SR level rather than the speech act level.

This system is capable of producing flexible CRs that adapt to the dialogue context, and this shows the value of the idea of modelling in a fine-grained way sources of CRs and of extending the concept of confidence scores. However, the system is only a first proof-of-concept, and we discuss possible improvements in the next section.

5 Conclusions and Further Work

We have presented a model of causes for requesting clarifications in dialogues. We classified these causes—

BR1 If two (or more) hypotheses are bridged via 'next' to same antecedent, closer is better.

BR3 Tomorrow should be referred to as 'tomorrow'.

RR1 If there are hypotheses where Plan-Corr has been inferred (non-monotonically) and some where other relations have been inferred, prefer these other hypotheses.

Figure 5: Some of the scoring rules

understanding problems in the widest sense—according to the level of processing on which they arise, and according to the severity of the problem. To make this precise, we related the multi-level models of communication of (Clark, 1996) and (Allwood, 1995) to the discourse semantics theory SDRT (Asher and Lascarides, 2003), and arrived at a fine-grained model of different understanding tasks which was motivated by analysing examples of CRs. We then proposed to extend the notion of *confidence score* from speech recognition to other kinds of processing (semantic and pragmatic), and sketched an implementation of this idea. We think that the resulting, relatively natural clarification behaviour shows that this idea of using 'pragmatic confidences' is promising.

However, the initial results also suggest that there is a lot of further work to be done. Firstly, it turned out during development of the system that setting the thresholds in the system manually in such a way that the desired behaviour was produced was rather hard (besides being *ad hoc*). We are currently exploring techniques to automatically learn the best settings from a corpus (this could perhaps be done along the lines of (Walker et al., 2000)). Secondly, the system we extended, RUDI, makes rather high demands on the quality of the data, being a system that relies on 'deep processing' at all stages. We are currently exploring ways of implementing the idea of using confidence values throughout in 'simpler', more realistic dialogue systems. This is a precondition for a thorough evaluation of the proposed clarification strategy, using 'real-world' criteria like user satisfaction and dialogue duration until task-completion.

With regard to the theoretical analysis of CRs, we are currently testing the coverage and accuracy of the model in a corpus study, and we are also working on a proper formalisation of the different classes of CR we proposed, in the framework of SDRT.

Acknowledgements

Parts of the work reported here are based on work done in collaboration with Alex Lascarides when the author was at the University of Edinburgh. Thanks are also due to Manfred Stede and Malte Gabsdil, who commented on drafts of the paper, and to the SIGdial reviewers for their helpful comments.

References

Jens Allwood. 1995. An activity based approach to pragmatics. Gothenburg Papers in Theoretical Linguistics 76, Göteborg University, Göteborg, Sweden.

Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press.

Herbert H. Clark. 1996. *Using Language*. Cambridge University Press, Cambridge.

Malte Gabsdil and Johan Bos. 2003. Combining acoustic confidence scores with deep semantic analysis for clarification dialogues. In *Proceedings of the 5th international workshop on computational semantics (IWCS-5)*, Tilburg.

Malte Gabsdil. 2003. Clarification in spoken dialogue systems. In *Proceedings of the 2003 AAAI Spring Symposium. Workshop on Natural Language Generation in Spoken and Written Dialogue*, Stanford, USA.

Jonathan Ginzburg and Robin Cooper. 2001. Resolving ellipsis in clarification. In *Proceedings of the 39th Meeting of the ACL*, Tolouse, France.

Herbert Paul Grice. 1989. *Studies in the Way of Words*. Harvard University Press, Cambridge, Massachusets.

Hans Kamp and Uwe Reyle. 1993. *From Discourse to Logic*. Kluwer, Dordrecht.

Staffan Larsson. 2003. Interactive communication management in an issue-based dialogue system. In *Proceedings of SemDial-7 (DiaBruck)*, Saarbrücken.

Matthew Purver, Jonathan Ginzburg, and Patrick Healey. 2001. On the means for clarification in dialogue. In *Proceedings of the 2nd SIGdial Workshop on Discourse and Dialogue*, Aalborg, Denmark.

R. San-Segundo, J.M. Montero, J. Ferreiros, R. Córdoba, and J.M. Pardo. 2001. Designing confirmation mechanisms and error recovery techniques in a railway information system for spanish. In *Proceedings of the 2nd SIGdial Workshop on Discourse and Dialogue*, Aalborg, Denmark.

David Schlangen, Alex Lascarides, and Ann Copestake. 2001. Resolving underspecification using discourse information. In *Proceedings of SemDial-5 (BiDialog)*, pages 79–93, Bielefeld.

David Schlangen. 2003. *A Coherence-Based Approach to the Interpretation of Non-Sentential Utterances in Dialogue*. Ph.D. thesis, School of Informatics, University of Edinburgh, Edinburgh, UK.

Marilyn Walker, Jerry Wright, and Irene Langkilde. 2000. Using natural language processing and discourse features to identify understanding errors in a spoken dialogue system. In *Proceedings of the 17th international conference on machine learning*.