Clarification in Spoken Dialogue Systems

Malte Gabsdil
Deptartment of Computational Linguistics
Saarland University
Germany
gabsdil@coli.uni-sb.de

Abstract

This paper argues for the use of more natural and flexible clarification questions in spoken dialogue systems. Several forms and aspects of clarification questions are discussed and exemplified by examples from human-human dialogue corpora. We show how clarification questions fit into the broader context of miscommunication and describe the types of information spoken dialogue systems need in order to generate various different clarification forms. Two kinds of dialogue systems are considered: Slot-based dialogue systems and dialogue systems with deep semantic analysis.

Introduction

Establishing mutual knowledge or *grounding* is a vital part of the communicative process (Allwood, Nivre, & Ahlsén 1992; Traum 1994; Clark 1996). But whereas managing grounding issues seems to be an almost effortless process for humans (often accomplished with simple feedback strategies or "backchannels"), it is a big challenge for spoken dialogue systems that have to deal with imperfect speech recognition. Some systems adopt a cautious grounding strategy, which requires the user to explicitly confirm all information provided to the system. The problem with cautious grounding is that it leads to very unnatural and inefficient dialogues. In extreme cases, systems pose much more clarification questions than task related questions, as shown in the following session with the GoDiS spoken dialogue system (Larsson & Ericsson 2002).

(1) GoDiS interaction
S: Where do you want to go?
U: From London to Paris on the third of October.
S: From London?
U: Yes.
S: To Paris?
U: Yes.
S: On the third of October?
U: Yes.

The opposite of cautious grounding is optimistic grounding, where the system accepts all user input without confirmation. The obvious drawback of this strategy is that it is prone to misunderstandings which are often difficult to recover from.

There are, of course, middle ways between optimistic and cautious grounding strategies. The Philips train timetable information system (Aust *et al.* 1995) was one of the first systems to use implicit confirmations, where the information provided by the user in the previous utterance is integrated into the next information request of the system (e.g. User: "I want to travel to Berlin." System: "When do you want to travel to Berlin?"). It is also possible to combine various different grounding strategies in a dialogue system. For example, a system can decide for each user utterance whether to accept (i.e. ground optimistically), confirm (i.e. ground cautiously, possibly using implicit confirmation), or reject it, based on the acoustic confidence score computed by the speech recognition engine. However, even combined grounding strategies often lead to unnatural system behavior. The main problem is that although systems exhibit different grounding strategies, they are very poor at generating the wide range of clarification forms found in human-human interactions. For instance, humans tend to clarify only the specific parts of an utterance they did not understand. Many current spoken dialogue systems are not able to do this – often they can only decide for whole utterances whether clarification is needed or not.

The aim of this paper, thus, is to explore the potential of more flexible clarification strategies for spoken dialogue systems based on insights from spoken dialogue data. In particular, we will look at how acoustic confidence scores (as returned by standard speech recognizers) can be used to generate different clarification forms.

Organizer

The paper is organized as follows. In the next section, we define the term clarification question and give examples of three key aspects that make them very efficient in human-human interactions: (i) partial clarification requests (i.e. clarification questions that only inquire about a particular sub-constituent of what was said), (ii) mention of explicit alternatives, and (iii) reformulation of information in relation to the task the user is performing. We then discuss the notion of clarification in the broader context of miscommunication. In particular, we will look at the three types of miscommunication introduced by (Hirst *et al.* 1994), namely misunderstanding, non-understanding and misconception. The subsequent sections work out what information spoken dialogue systems need in order to generate more flexible clarification requests. First, we consider slot-based dialogue systems which are widely used for simple task-oriented dialogues like travel information and booking. We then turn to systems in more complicated domains that require deeper semantic analysis and knowledge representation. The final section summarizes and concludes the paper.

Clarification Questions

We define the term *clarification question* or *clarification request* in a very narrow sense. Intuitively, a speaker asks a clarification question only when he did not (fully) understand or is uncertain about what the previous speaker said or meant with an utterance. Under this view clarification questions should be assigned the backward looking function *signal non-understanding* in the DRI classification scheme of dialogue acts (Core *et al.* 1998). Within the DRI scheme, clarification questions can therefore be analyzed as a subtype of check-questions (which also appear in the MapTask annotation manual (Carletta *et al.* 1996)).

A good test for identifying clarification questions is that they cannot be preceded by explicit acknowledgments, as indicated by the following examples.

(2) Clarification question (reprise fragment)

Lara:	There's only two people in the class.
a) Matthew:	Two people?
b) Matthew:	(??) OK, I see. Two people?

(BNC, taken from (Purver, Ginzburg, & Healey 2001))

(3) Other check question

U:	How long will it take to Avon?
a) S:	With engine E one?
b) S:	OK, let's see. With engine E one?

(TRAINS corpus, (Allen *et al.* 1995))

(2b) is odd because Matthew first acknowledges Lara's turn and then asks whether he correctly understood a specific part of it. This could be a felicitous move but requires a very marked intonation or a long pause which would induce some kind of "backtracking" effect. On the other hand, (3b) is fine: S needs some additional information to be able to answer U's initial question.

This definition of clarification questions is essentially equivalent to the one adopted in (Purver, Ginzburg, & Healey 2001; Ginzburg & Cooper 2001).[1]

Partial Clarification Questions

In a study of a 150.000 word dialogue sub-corpus of the BNC, (Purver, Ginzburg, & Healey 2001) identified several different syntactic forms of clarification questions. An interesting fact about their study is that about 45% of the identified clarifications are fragmentary/elliptical (category "reprise fragments") or result from replacing a constituent in the to-be-clarified utterance with a wh-word (categories "reprise sluices" and "wh-substituted reprises"). An example of a reprise fragment clarification is given in (2) above. (4) and (5) give examples of reprise sluices and wh-substituted reprises, respectively (both are taken from (Purver, Ginzburg, & Healey 2001)).

(4) Reprise sluice

Sarah:	Leon, Leon, sorry she's taken
Leon:	Who?
Sarah:	Cath Long, she's spoken for.

(5) Wh-substituted reprise

Unknown:	He's anal retentive, that's what he is.
Kath:	He's what?
Unknown:	Anal retentive.

The high frequency of these kinds of examples indicates that people often only clarify the parts or sub-constituents of utterances they did not understand. This is a very efficient way to reveal to the other dialogue participant what exactly caused the understanding problem. Partial clarification questions are therefore much more informative than what (Purver, Ginzburg, & Healey 2001) call "non-reprise" and "conventional" clarification questions (e.g. "What did you say?" and "Pardon?") that reject utterances as a whole.

Alternative Clarification Questions

Another interesting form of clarification questions is the explicit mention of the alternative interpretations an utterance gives rise to. The following two examples are both taken from everyday conversations.

(6) Acoustic ambiguity

A:	Denilson scored.
B:	Denilson or Edilson?

(7) Referential ambiguity

A:	Did you hear? George Bush is in hospital.
B:	Junior or Senior?

Semantically, alternative questions can be analyzed as multiple yes/no-questions. Answering one of them positively conversationally implicates that all other options are excluded (see e.g. (Levinson 1983) for a discussion of scalar implicatures). Speakers can therefore make use of alternative questions to encode more information than in a single polar question. However, there seems to be a natural limit to the possible number of choices within an alternative question. Most frequent are alternative questions which offer two options, three options are already rare, and more than three are the exception.

Task-level Reformulations

Instead of echoing (parts of) a previous utterance, people also make use of task-level reformulations for clarification. Task-level reformulations are typically used to confirm more complex actions, for example in instruction giving dialogues. The initiator of the clarification request reformulates

[1] An anonymous reviewer suggested that clarification questions might also be distinguished from other questions by intonation. For research in this direction on the MapTask corpus see e.g. (Hastie, Poesio, & Isard 2002). It has also been shown that so called echo questions have a distinguished intonation contour (see e.g. (Hirst & Cristo 1998)).

the effects of an utterance, thereby showing his subjective understanding to the other dialogue participant. (8) gives an example from the MapTask corpus (Anderson *et al.* 1991). The instruction in (9) is taken from the IBL corpus (Lauria *et al.* 2002). Since the IBL corpus only contains route instructions, the clarification request in (9) was made-up by the author.

(8) MapTask example

G: No, I want you to go up the left hand side of it towards green bay and make it a slightly diagonal line, towards, em sloping to the right.

F: So you want me to go above the carpenter?

(9) IBL example

A: You turn the second road on your left hand side.

B: You mean Marchmont Road?

Reformulations are particularly useful in task-oriented dialogues. Rephrasing an utterance with respect to the task confirms its practical implications in a given context. What distinguishes task-level reformulations from partial and alternative forms of clarification is that they are in principle independent of the actual surface form of the to-be-clarified utterance, which makes them interesting candidates for template based generation. We will later see that task-level reformulations can easily be incorporated into slot-based dialogue systems.

Miscommunication and Levels of Communication

It is useful to view clarifications in the broader context of miscommunication. Following (Hirst *et al.* 1994), miscommunication can be partitioned into three different types: *Misunderstanding*, *non-understanding*, and *misconception*. We discuss these three types in more detail below.

Misunderstanding

Misunderstandings differ from non-understandings and misconceptions in that they are not detected right away. Rather, the hearer obtains an interpretation that "she believes is complete and correct but which is, however, not the one the speaker intended her to obtain" (Hirst *et al.* 1994). We do not address misunderstandings in this paper because they typically lead to corrections instead of clarifications (i.e. information being revised instead of confirmed).

Non-Understanding

The second type of miscommunication, non-understanding, occurs whenever the hearer "fails to obtain any interpretation at all or obtains more than one interpretation, with no way to choose among them" (Hirst *et al.* 1994). We want to propose that non-understanding should additionally include cases where the hearer is uncertain about what was said. The difference between ambiguous and uncertain understanding is that the hearer has to choose from a set of explicit alternatives in the former whereas there is only one interpretation available in the latter. In our view, this difference leads to different forms of clarifications: Uncertain interpretations can coarsely be associated with single polar questions whereas ambiguous understanding is more likely to result in alternative questions or wh-questions (when the number of alternatives exceeds a certain limit).

Furthermore, non-understanding in general can occur on several different communicative levels, ranging from establishing contact among the dialogue partners to the intended meaning or function of the utterance in context. (Clark 1996) argues for four basic levels of communication in a framework that views interaction as a joint activity of the dialogue participants. Clark's four levels are execution/attendance, presentation/identification, signal/recognition, and proposal/consideration. On the lowest level, dialogue participants establish a communication channel, which is then used to present and identify signals on level two. On level three, these signals are interpreted before their communicative function is evaluated on the proposal/consideration level. The framework of joint actions requires that dialogue participants coordinate their individual actions on all of those levels.

When we combine the sources of non-understanding with Clark's four levels of communication, we end up with the following picture:

proposal/ consideration		Is this a question?	
signal/ recognition	Where's Avon?	Avon, Colorado?	Junior or Senior?
presentation/ identification	Pardon?	To Avon?	Denilson or Edilson?
execution/ attendance		Are you talking to me?	
	no interpr.	uncertain interpr.	ambiguous interpr.

Table 1: Sources and levels of non-understanding

Table 1 pairs sources of non-understanding (horizontally) with Clark's levels of communication (vertically) and gives some examples of clarifications that may arise when these two dimensions are combined. Table 1 should only be seen as a coarse-grained classification of clarifications. In particular, we suspect that the signal/recognition level (which comprises several different aspects of the understanding process) might need more refinement. On the other hand, this rather broad picture fits well with the two major readings for clarifications proposed by (Ginzburg & Cooper 2001). Their "clausal reading" can be related to the presentation/identification level and their "constituent reading" to the signal/recognition level. Clausal readings take as a basis the content of the utterance being clarified and can therefore roughly be paraphrased as "Are you asking/asserting that X?" or "For which X are you asking/asserting that Y?". Constituent readings, on the other hand, inquire about the content of a constituent of the previous utterance. They can roughly be paraphrased as "What/who is X?" or "What/who do you mean by X?" (for details, see (Ginzburg & Cooper 2001)).

Our discussion so far might suggest that every clarification request can be unambiguously assigned a specific position in Table 1. But this is not the case. (Ginzburg & Cooper 2001) have shown that clarifications can be ambiguous with respect to their readings. Also, (Derriks & Willems 1998) argue, based on a study of French train-timetable information dialogues, that the conventional "pardon?" (sorry) is highly ambiguous in the sense that it can indicate different communication problems. The following three examples from their paper are interesting for our discussion of clarification questions (we only quote the English translations of their data).

(10) Example 1 from (Derriks & Willems 1998)

C62: I have forgotten where you have to go for the tickets
O61: **sorry**
C63: where do you have to go for the tickets
O62: ah, you can reserve by telephone if you like. You can reserve your/your tickets

(11) Example 2 from (Derriks & Willems 1998)

O24: well
C23: it's perhaps for that reason that she hasn't done it for me
O25: **sorry?**
O24: it's perhaps for that reason that she hasn't done it for me

(12) Example 3 from (Derriks & Willems 1998)

C2: good day. Tell me madam, eh I would like to know at what time I can take a train and (know) which train goes to the city of Saint-Hilaire de Harcouèt
O2: **sorry**, Saint-Hilaire du Harcouèt
C3: yes

According to their analysis, (10) "signals an uncertainty in relation to the perception or comprehension in the previous utterance", (11) "signals global incomprehension", and (12) constitutes "a mark of astonishment or uncertainty". Examples (10) and (11) can therefore be associated with our categories of uncertain interpretation (at the presentation/identification level) and no interpretation (at the presentation/identification or the signal/recognition level) respectively. Example (12) can either be classified as an uncertain understanding at the presentation/identification level or a non-understanding at the signal/recognition level. Yet another possibility is that it is an instance of a different kind of clarification that is related to the third (and final) type of misunderstanding, namely misconception.

Misconception

Misconceptions occur when the "hearer's most likely interpretation suggests that beliefs about the world are unexpectedly out of alignment with the speaker's" (Hirst *et al.* 1994). Misconceptions often lead to clarifications that resemble presentation/identification level clarifications but have a very distinguished "surprise" intonation. (13) gives an example where Susan's beliefs about proper counting are in conflict with Steven's.

(13) Misconception

Steven: One, two, three ((pause)) four, five, six, ((pause)) eleven, eight, nine, ten.
Susan: "E*le*ven"? – eight, nine, ten?
Steven: Eleven, eight, nine, ten.
Susan: "E*le*ven"?
Steven: Seven, eight, nine, ten.
Susan: That's better.

(Jefferson 1972)

Clarifications in response to misconceptions usually convey extra-linguistic information like surprise or astonishment. The example in (13) might even be analyzed as expressing disbelief or a polite form of correction.

In the next two sections we look in more detail at how spoken dialogue systems might be able to generate clarification questions on the presentation/identification level, based on the output of standard speech recognizers. We consider two different types of systems: first, slot-based dialogue systems and second dialogue systems with deep semantic analysis.

Slot-based Dialogue Systems

Many current spoken dialogue systems employ the technique of *slot-filling* to recover semantic information from the speech recognition component. In this section, we are considering the simplest form of slot-based systems, namely non-recursive slot-based systems. These systems allow only one atomic value for each slot. It is not possible to define recursive or nested data-structures which again hold a record of slot-value pairs as a value.[2] Indeed, many off-the-shelf recognizers are tailored to this specific kind of application. A case in point is the commercial Nuance speech recognition and dialogue management package (`http://www.nuance.com`) or Scansoft's ASR1600 SDK (former Lernout and Hauspie, `http://www.scansoft.com`). To be fair, we should note that the Nuance software also allows nested record structures and list values.

Slots and Confidence Scores

Speech recognizers typically output an N-best list of recognition hypotheses decorated with confidence scores. These confidences can then be used by the dialogue system to decide how to proceed in the dialogue (e.g. whether clarification is needed or not). Confidence scores are computed for whole utterances, individual words, and slot values, the latter based on the confidences of the words that triggered the value of specific slots. For example, the Nuance system is able to return several (i.e. N-best) structures of the following form for the utterance "Withdraw fifteen hundred dollars from savings":

(14) Example Nuance output

[2] For convenience, we will stick to the term "slot-based systems" instead of the lengthy "non-recursive slot-based systems"

Slot	Value	Confidence
action	withdraw	49
amount	$1500.00	60
account	savings	61
Overall confidence:		56

The easiest way to decide whether an utterance needs clarification is to consider only the overall confidence score returned by the recognizer. A common practice is to use two fixed thresholds to decide whether an input should be rejected entirely ("Sorry, I didn't understand"), needs confirmation ("Withdraw 1500$ from savings, is this correct?"), or to accept it without clarification.

Partial Clarification Questions

The information in (14) can also be used to generate partial clarification requests based on the individual slot confidences. For example, a system might decide that the confidence scores for the amount and account slots are sufficiently high to accept but that it needs to clarify the action. This can be done by using a task-level template of the form "Do you want to WITHDRAW *$amount* Dollars from your *$account* account?".[3] Note that this template only accidentally repeats the surface form of the user utterance. The same template would be used for all input utterances that lead to similar slot representations (e.g. "Take fifteen hundred out of savings" or "Give me one thousand five hundred dollars from my savings account").

Alternative Clarification Questions

In case the possible values for a slot are limited, the system has all necessary information to decide whether an alternative clarification request can be generated. Again, using a template the system might ask a question like "Do you want to WITHDRAW or DEPOSIT *$amount* Dollars?". Furthermore, since it is known what the slots "mean" (i.e. slots implicitly encode sortal information), the system can in principle generate clarification questions like "HOW much money do you want to withdraw?" replacing the amount value with an appropriate question phrase. Slot-based dialogue systems can additionally make use of the N-best results returned by the recognizer to generate alternative clarifications (e.g. in the case where the same slot is filled with different values in different results). Finally, we can imagine simple forms of clarification questions dealing with referential ambiguities. For example, a database lookup for airports in London in a travel agency domain may return several results which can then be assembled into a clarification like "Do you mean London Heathrow or London Stansted?".

Limitations

The potential of generating clarifications for slot based systems is restricted by the limited amount of linguistic information these system have access to. If only slots and values are considered, the natural way to generate clarifications is to use pre-canned texts or sentence templates that amount to simple task-level reformulations. In example (14), one can imagine different clarification templates that are filled and used depending on which values for which slots are available. However, based on slot-value pairs alone, there is no possibility to generate a variety of clarification forms discussed by (Purver, Ginzburg, & Healey 2001) that depend on syntactic parallelism constraints. Even the apparently simple case of "literal reprises", where (parts/constituents of) the to-be-clarified sentence are repeated cannot be generated without information about the surface form and the basic syntactic constituent structure of the input utterance. An interesting line of research for making slot-based systems more flexible with respect to their clarification behavior would therefore be to further analyze the part of an utterance that lead to a specific slot being filled (e.g. its surface and syntactic form, as well as the confidences assigned to the words that make up the slot). Assume, for example, that the amount slot in (14) above is composed from the three words "fifteen", "hundred", and "dollars". If the confidence for "fifteen" is much lower than the ones for "hundred" and "dollars", the system might clarify this with the query "FIFTEEN hundred dollars?", stressing the first word of the constituent.

[3]Words written in capital letters indicate the "intonation center" of an utterance.

Dialogue Systems with Deep Semantic Analysis

The slot-filling paradigm is successful in several applications such as travel service and booking, reservation confirmation, price and availability information, and call routing. But for speech applications that demand a deeper analysis of semantic information, slot-filling approaches soon become a burden for serious semantic interpretation. Dialogue systems in the area of home automation show the need for a proper treatment of quantification and negation (Bos & Oka 2002; Quesada *et al.* 2001). Speech-controlled mobile robots often have such a rich scenario of primitive actions that slot-filling is simply unfeasible (Lauria *et al.* 2001). Typical problems in more complex scenarios are negation, coordination phenomena, and quantified NPs as exemplified in (15) and (16).

(15) Home automation
  A: Switch off all lights except the one in the kitchen.

(16) Route instruction
  A: First go straight and then turn left at Tesco.

Systems operating in these domains employ richer semantic representations and apply ambiguity resolution in a second stage for sufficient understanding of speaker's utterances. It is in general not possible to represent the contents of more complex utterances in (non-recursive) slot-based systems appropriately.

Abstract Representations

Generating clarifications in systems with deep semantic processing is a much harder task than in slot-based systems. The basic problem is that these systems do not have individual slot-confidences to start with. It is a non-trivial problem to retain the link between acoustic confidences and parts of more abstract (quasi-)logical formulas. The easiest way out would be to "back-off" to the confidence score assigned to the whole utterance, but then we are faced with the same problems as described in the previous section for slot-based systems: we can only decide on a very general level whether to accept, confirm, or reject an utterance, which brings us back to conventional clarifications like "Sorry, what did you say?" or "Pardon?".

Already the mere confirmation of what the user said cannot, in general, be achieved in systems with deep analysis by filling in templates with given slot values. Imagine, for example, a system that uses formulas of first-order logic as semantic representations. To clarify a user utterance as a whole, such a system would have to generate a clarification question from its logical representation. To give a concrete example let us assume that the logical representation of (15) is (17), where the "!" indicates that it should be interpreted as a command.

(17) Logical representation of (15)

$!\forall(x)[\text{light}(x) \wedge \neg\text{in}(x,\text{kitchen}) \rightarrow \text{switch\_off}(x)]$

Obviously, we need a more powerful generation component to formulate a polar clarification request like "Do you want me to ..." from this abstract representation.

Partial Clarification Questions

The generation of partial or alternative clarification requests requires even more efforts. In order to generate partial clarifications, systems first have to decide which parts of an utterance should be clarified. This means that systems have to make use of the confidence scores the recognizer assigned to the individual words of the input utterance. (Gabsdil & Bos 2003) have proposed a first method of how individual word confidences can be integrated into subsequent steps of linguistic processing (i.e. parsing and semantic construction). In this approach, confidence scores are attached to labels that identify parts of logical formulas in an underspecified semantic formalism.

Another problem for systems with deep semantic analysis is that they cannot rely on information that is implicitly present in slot-based approaches. Suppose, for example, that a system has identified a certain constituent in the input utterance that should be clarified with a wh-substituted reprise or a reprise sluice. Whereas in slot-based systems we can associate with each slot a certain semantic type or even a specific wh-phrase, this is generally not possible when deep semantic representations are used. Systems need access to ontological knowledge and information about the selectional restrictions of verbs to infer which wh-word or phrase must be used in different situations. Let us exemplify this with the following example from the IBL corpus, where the numbers in parenthesis indicate the individual word confidence scores assigned by a speech recognizer.

(18) IBL route instruction example

R: again(66) you(60) walk(49) straight(28) ahead(58)

If the system was to generate a clarification like "Sorry, which direction?" from a (quasi-)logical representation of the utterance, it has to know the argument frame of the verb "walk". We can, of course, try to generate a clarification question solely on the basis of the individual words and their confidences (i.e. without sortal information). For the example in (18), the system might be able to come up with the polar question "STRAIGHT?"; if it had additional syntactic information, this would allow to clarify the sub-constituent the low-confidence word "straight" is part of, (i.e. "STRAIGHT ahead?"). Note, however, that the question "Sorry, which direction?" is unambiguous whereas the plain "STRAIGHT?" also has a misconception reading ("Straight, how? There is a wall" or "Straight, why? This is where we came from") and might, at least technically, also allow for an interpretation on the signal level (i.e. "What do you mean by straight?").

A more natural (but made-up) example is given in (19), where B's clarification question can be either paraphrased as "Did you say that *Peter* called?", "*Peter*? I thought he is on vacation", or "Who is Peter?".

(19) Ambiguous clarification question

A: Peter called.
B: Peter?

The different interpretations could probably be distinguished by intonation. Intuitively, the constituent reading ("Who is Peter") has a rising intonation whereas the clausal reading ("Are you asserting that *Peter* called?") has a fall followed by a rise.[4]

Alternative Clarification Questions

The main problem in generating alternative clarification requests in systems with deep semantic processing is that we do not have fixed alternatives based, for example, on two or more different slot representations. Given the N-best recognizer hypotheses, the system has to figure out which information distinguishes the different interpretations (which might be an undecidable task). (Rosé 1997) has pioneered an approach that is able to identify features that distinguish competing interpretations in (deeply nested) frame-based semantic representations. She was also able to construct yes/no clarification questions on the task level for some representations which exclude one or more other possible interpretations. (20) gives an example.

(20) Rosé's example (Rosé 1997)

Input: What about any time but the ten to twelve slot on Tuesday the thirtieth?

Recognition$_1$: How about from ten o'clock till twelve o'clock Tuesday the thirtieth any time?

Recognition$_2$: From ten o'clock till Tuesday the thirtieth any time?

DRose: Are you suggesting that Tuesday November the thirtieth from 10 a.m. to 12 a.m. is a good time to meet?

[4] An anonymous reviewer suggested work by Mariët Theune (see e.g. (Theune 2000)) as an example for a system that makes use of different accentuation markings in the speech output.

The system comes up with two different interpretations for the input sentence (based on the two recognition results). Both of them contain a suggestion for a possible meeting time. Rosé is able to spot that both interpretations are suggestions and only differ with respect to the proposed time-span.[5] Given this information, the system is able to generate a task-level clarification question that tries to confirm the first interpretation. The example might be a little bit misleading because if the user answered the clarification question with "No", the system would chose the second interpretation which is even further away from what the speaker intended. However, it should be clear that the clarification question in (20) distinguishes the two possible interpretations based on their different time spans.

Another option to detect possible alternatives is to use a speech lattice as input for the linguistic processing modules. A practical problem with this approach is that off-the-shelf recognizers simply do not output lattice structures. Another problem is that speech lattices may encode several hundred or thousands of different hypotheses. Even if these alternatives could be pruned down to a manageable size, there might still be the problem that similar spans in the lattice have different syntactic categories which cannot be put together into an alternative question (e.g. consider the two different interpretations in (20) above).

Task-level Reformulations

We have already briefly mentioned the generation of task-level reformulations based on the semantic representation of the user's input utterance. Instead of rephrasing what was said, it might also be possible to generate reformulations that focus on the implications a given interpretation has on the task (e.g. by looking at the effects of action operators that are triggered by an interpretation). (Gabsdil, Koller, & Striegnitz 2002) generate descriptions of locations and objects based on actions a player performs in a text-based computer game. However, it is an open question whether their approach can be generalized to dialogue applications.

Conclusions

Clarification questions play a significant role for ensuring mutual understanding in spoken dialogue systems. Imperfect speech recognition requires dialogue systems to confirm user input much more frequently than in human-human interaction. It is therefore necessary to equip dialogue systems with clarification strategies that are as flexible and natural as possible. In this paper, we have discussed the notion of clarification in general and have looked at three common aspects of clarifications in more detail: First, partial clarification questions that only clarify sub-constituents of the to-be-clarified utterance. Second, alternative clarification questions that offer different interpretations, and third reformulations that relate an utterance to the effects it has on the task-level. We also looked at clarifications from the more general viewpoint of miscommunication, where they can be seen as typical responses to non-understandings and misconceptions. We proposed to extend the standard subdivision of non-understanding with the new category uncertain understanding and argued that uncertain understanding, like ambiguous understanding and (total) non-understanding leads to characteristic forms of clarification questions.

In slot-based dialogue systems, the confidence scores that are assigned to individual slot values are a good starting point for generating clarification requests. Slot-based systems lend themselves to pre-canned or template based generation. However, looking only at slot values excludes a wide range of clarification forms. We argued that combining slot-based systems with a more fine-grained analysis of the strings that trigger specific slot values might result in more interesting and flexible clarification questions. Systems with deep semantic processing face much harder problems in generating clarification requests. First of all, recognizer confidences have to be integrated into the linguistic processing components. Furthermore, interfaces to knowledge sources like ontologies or selectional restrictions of verbs are necessary to generate wh-substituted clarifications or reprise sluices. Generation of alternative clarifications for systems with deep semantic processing is particularly hard, because it presupposes that alternatives which distinguish two or more interpretations can be computed. Task-level reformulations, on the other hand, might benefit from systems that have access to effects of action operators or other ways to compute task-level implications.

The ultimate goal of using more flexible clarification strategies in spoken dialogue system obviously is to improve human-machine interaction. Clarification is a part of communication which is and will be very prominent in spoken dialogue systems, especially as long as speech recognizers remain fallible. Given the high number of clarification questions needed in spoken dialogue systems, we suspect that the generation of more complex clarification forms on the system side will improve human-machine interaction and therefore lead to better system usability and user satisfaction.

Acknowledgements

I would like to thank two anonymous reviewers for valuable comments on this paper. Johan Bos and Ivana Kruijff-Korbayová read earlier versions of the paper and Martine Grice pointed me to literature on question intonation. I also want to thank Nuance Communications Inc. for making available their recognition software for research purposes.

References

Allen, J. F.; Schubert, L. K.; Ferguson, G. M.; Heeman, P. A.; Hwang, C. H.; Kato, T.; Light, M. N.; Martin, N. G.; Miller, B. W.; Poesio, M.; and Traum, D. R. 1995. The TRAINS project: A case study in building a conversational planning agent. *Journal of Experimental and Theoretical AI* 7:7–48.

Allwood, J.; Nivre, J.; and Ahlsén, E. 1992. On the Semantics and Pragmatics of Linguistic Feedback. *Journal of Semantics* 9:1–26.

[5]Time-spans here are not simply encoded as two slots representing *begin* and *end*. They are complex structures that allow to encode a variety of different time expressions including "in the morning", "after 2 pm", "Monday or Tuesday", etc.

Anderson, A. H.; Bader, M.; Bard, E. G.; Boyle, E.; Doherty, G.; Garrod, S.; Isard, S.; Kowtko, J.; McAllister, J.; Miller, J.; Sotillo, C.; Thompson, H. S.; and Weinert, R. 1991. The HCRC Map Task Corpus. *Language and Speech* 34(4):351–366.

Aust, H.; Oerder, M.; Seide, F.; and Steinbiss, V. 1995. The Philips automatic train timetable information system. *Speech Communication* 17:249–262.

Bos, J., and Oka, T. 2002. An Inference-based Approach to Dialogue System Design. In *Proceedings of Coling 2002*.

Carletta, J.; Isard, A.; Isard, S.; Kowtko, J.; Doherty-Sneddon, G.; and Anderson, A. 1996. HCRC Dialogue Structure Coding Manual. Technical Report HCRC/TR-82, Human Communication Research Centre, University of Edinburgh.

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

Core, M.; Ishizaki, M.; Moore, J.; Nakatani, C.; Reithinger, N.; Traum, D.; and Tutiya, S. 1998. The Report of the Third Workshop of the Discourse R esource Initiative. Chiba University and Kazusa Academia Hall.

Derriks, B., and Willems, D. 1998. Negative feedback in information dialogues: identification, classification and problem solving procedures. *International Journal of Human-Computer Studies* 48(4):577–604.

Gabsdil, M., and Bos, J. 2003. Combining Acoustic Confidence Scores with Deep Semantic Analysis for Clarification Dialogues. In *Proceedings of the Fifth International Workshop on Computational Semantics (IWCS-5)*.

Gabsdil, M.; Koller, A.; and Striegnitz, K. 2002. Natural Language and Inference in a Computer Game. In *Proceedings of Coling 2002*.

Ginzburg, J., and Cooper, R. 2001. Resolving Ellipsis in Clarification. In *Proceeding of ACL-01*.

Hastie, H. W.; Poesio, M.; and Isard, S. 2002. Automatically predicting dialogue structure using prosodic features. *Speech Communication* 36:63–79.

Hirst, D., and Cristo, A. D., eds. 1998. *Intonation Systems. A Survey of Twenty Languages*. Cambridge University Press.

Hirst, G.; McRoy, S.; Heeman, P.; Edmonds, P.; and Horton, D. 1994. Repairing Conversational Misunderstandings and Non-Understandings. *Speech Communication* 15:213–230.

Jefferson, G. 1972. Side Sequences. In Sudnow, D., ed., *Studies in Social Interaction*. New York: Academic Press. 295–330.

Larsson, S., and Ericsson, S. 2002. GoDiS – Issue-Based Dialogue Management in a Multi-Domain, Multi-Language Dialogue System. In Smith, R., ed., *Demonstration Abstracts, ACL-02*.

Lauria, S.; Bugmann, G.; Kyriacou, T.; Bos, J.; and Klein, E. 2001. Training Personal Robots Using Natural Language Instruction. *IEEE Intelligent Systems* 38–45.

Lauria, S.; Bugmann, G.; Kyriacou, T.; and Klein, E. 2002. Mobile Robot Programming Using Natural Language. *Robotics and Autonomous Systems* 38(3-4):171–181.

Levinson, S. C. 1983. *Pragmatics*. Cambridge University Press.

Purver, M.; Ginzburg, J.; and Healey, P. 2001. On the Means for Clarification in Dialogue. In *Proceedings of the 2nd ACL SIGdial Workshop on Discourse and Dialogue*, 116–125.

Quesada, J. F.; Garcia, F.; Sena, E.; Bernal, J. A.; and Amores, G. 2001. Dialogue management in a home machine environment: Linguistic components over an agent architecture. In *Proceedings of the 17th SEPLN*, 89–98.

Rosé, C. P. 1997. *Robust Interactive Dialogue Interpretation*. Ph.D. Dissertation, School of Computer Science, Carnegie Mellon University.

Theune, M. 2000. *From data to speech: language generation in context*. Ph.D. Dissertation, Eindhoven University of Technology.

Traum, D. 1994. *A Computational Theory of Grounding in Natural Language Conversation*. Ph.D. Dissertation, Department of Computer Science, University of Rochester.