Clarification Questions with Feedback

Svetlana Stoyanchev, Alex Liu, Julia Hirschberg

Computer Science, Columbia University, New York, NY, USA

sstoyanchev@cs.columbia.edu, aliu945@gmail.com, julia@cs.columbia.edu

Abstract

In this paper, we investigate how people construct clarification questions. Our goal is to develop similar strategies for handling errors in automatic spoken dialogue systems in order to make error recovery strategies more efficient. Using a crowd-sourcing tool [7], we collect a dataset of user responses to clarification questions when presented with sentences in which some words are missing. We find that, in over 60% of cases, users choose to continue the conversation *without* asking a clarification question. However, when users *do* ask a question, our findings support earlier research showing that users are more likely to ask a *targeted* clarification question than a generic question. Using the dataset we have collected, we are exploring machine learning approaches for determining which system responses are most appropriate in different contexts and developing strategies for constructing clarification questions.[1]

Index Terms: clarification, question

1. Introduction

1.1. Clarifications in Human Dialogue

Clarification questions are common in human-human dialogue. They help dialogue participants maintain dialogue flow and resolve misunderstandings. A clarification question may be asked by a listener who fails to hear or understand part of an utterance. Requesting information is not the only role of a clarification question. It also helps ground communication by providing feedback indicating which information is known and understood.

In the following example [5], Speaker B has failed to hear the word *toast* and so constructs a clarification question using a portion of the correctly understood utterance — the word *some* — to query the portion of the utterance B has failed to understand:

A: Can I have **some** *toast* please?
B: Some?
A: Toast.

Such *targeted clarification questions* signal the location of the recogniton error to hearer. In this case, Speaker A then is thus able to respond with a minimal answer to the question — filling in only the missing information.

Human speakers employ diverse clarification strategies in dialogue. Examining human clarification strategies, Purver [5] distinguishes two types of clarification questions: *reprise* and *non-reprise* questions. Reprise questions are questions like B's query above, in which a portion of the interlocutor's utterance believed to have been recognized *correctly* is repeated as context for the portion believed to have been misrecognized or simply unheard. Non-reprise questions are simply generic requests for a repeat or rephrase of a previous utterance, such as *What did you say?* or *Please repeat*. Such questions do **not** include contextual information from the previous utterance. Reprise clarification questions, on the other hand, ask a *targeted* question about the part of an utterance that was misheard or misunderstood, using portions of the misunderstood utterance which are thought to be correctly recognized.

In human-human dialogues, reprise clarifications are much more common than non-reprise questions, which explicitly signal an error without providing information about its location. However, spoken dialogue systems predominantly use *non*-reprise strategies to indicate recognition errors to their users, typically requesting that the user repeat or rephrase their utterance [2]. Constructing non-reprise questions is significantly simpler than creating reprise questions and can easily be hard-coded in the system, since they do not include contextual information. However, to construct a reprise clarification question, a system must first determine which part of an utterance it believes contains an error. It must then construct an appropriate question based upon information in the correctly recognized part of an utterance.

In this paper we describe the collection of a corpus of clarification questions from Mechanical Turk [7] workers who were asked to indicate how they would respond to an utterance containing some unknown words. Such utterances were created from a set of misrecognized utterances in which blanks were substituted for recognition errors. We describe these annotators' recovery strategies, including the type of question asked or request made to recover missing information from an utterance. Our ultimate goal is to learn the relationship between clarification strategies and features of misrecognized utterances in order to develop automated methods for developing better

[1]This work was partially funded by DARPA HR0011-12- C-0016 as a Columbia University subcontract to SRI International.

error recovery strategies in spoken dialogue systems. We are currently developing such a process for a speech-to-speech (S2S) translation system in which the Dialogue Manager can query users about hypothesized misrecognitions, out-of-vocabulary (OOV) items, and translation errors before a translation is presented to the interlocutor.

1.2. Clarification in Speech-to-Speech Translation Systems

In a S2S translation system, two speakers communicate orally in two different languages through two ASR systems and two Machine Translation (MT) systems. Such a system takes speech in one language as input, recognizes it using an Automatic Speech Recognition (ASR) system, translates the recognized input into text in another language, and finally produces synthetic speech output from the translation for the conversational partner. In the S2S application we target, speakers converse freely about topics which may be pre-specified in very general terms. When an ASR is hypothesized in a speaker utterancer, the clarification component of the system seeks to clarify errors with the speaker before passing a corrected ASR transcription on to the MT component. In this way, the clarification component attempts to intercept speech recognition errors early in the dialogue to avoid translating poorly recognized utterances. In parallel research we have also developed a method for localized ASR error detection in the output of the speech recognizer of an S2S translation system.

The ability to produce reprise clarification questions in S2S translation systems is especially important. While in a form-filling dialogue system, clarification questions can be designed around a set of specific domain concepts, open-domain systems such information is not available. For example, if a user of a closed domain system, such as an airline reservation system, mentions a departure location which the system misrecognizes, the system may construct a predefined clarification question *Leaving from where?*. However an open-domain translation system must accept input on a variety of topics and cannot rely upon users mentioning a particular set of domain concepts. Reprise clarification questions constructed by such systems must be generated by the system dynamically. In our experiment, we collect questions for English utterances containing errors from an open-domain S2S translation system. Our motivation is to develop a reprise clarification strategy containing feedback and grounding information. We hypothesize that a system capable of asking clarification questions that are more similar to the types of questions that humans ask will be more natural and lead to more efficient error recovery.

In Section 2, we describe previous research on user responses to errors in spoken dialogue. In Section 3, we describe the data collection experiment and analyze our results. We conclude in Section 4 with our plan for the use of the described dataset for learning strategies in a dialogue system.

2. Related work

A number of researchers in spoken dialogue have studied user responses to errors in dialogue. For example, Skantze [6] collected and analyzed user responses to speech recognition errors in a direction-giving domain in Swedish, using a used speech a recognizer to corrupt human-human speech communication in one direction. Williams and Young [9] performed a Wizard-of-Oz study in a tourist information dialogue system in which recognition errors were systematically controlled. Koulouri and Lauria [4] performed another Wizard-of-Oz study in a human-robot instructions domain with the "wizard" playing a role of a robot with restricted communication capabilities. In all of these studies, results indicate that, when subjects encounter speech recognition problems, they tend to ask task-related questions providing feedback to the other speaker and confirming their hypothesis about the situation. These studies also find that speakers rarely give direct a indication of misunderstanding to the system that they have misunderstood, irrespective of the system's word-error-rate. Williams and Yaung's findings suggest that, at moderate speech recognition levels, asking task-related questions appears to be a more successful strategy for recovering from error tthan direct signaling of the error itself.

In our study, we collect a (text) corpus of human responses to missing information in ASR transcriptions. We will use this corpus in future to improve our dialogue clarification strategy by automatically creating targeted reprise clarification questions in response to errors in an open-domain S2S translation system. However, we believe this strategy will also be relevant to other open-domain spoken dialogue systems.

3. Experiment

3.1. Dataset

We perform our experiments on data from SRI's *Iraq-Comm* speech-to-speech translation system [1]. The data were collected by NIST during seven months of evaluation exercises performed between 2005 and 2008 [8]. The corpus contains acted dialogues between English and Arabic speakers. Table 1 shows a sample dialogue from the dataset, with correct English translations for the Arabic utterances. The dataset is manually transcribed. We tag the manual transcript of the dataset with part-of-speech (POS) tags using Stanford POS tagger [3]. We identify POS tags of misrecognized words by aligning the ASR output with the transcript.

In our data collection, we use 475 English utterances from the dataset.[2] Each utterance we present to an anno-

[2]This is an ongoing study and we are continuing to collect more

English:	good morning
Arabic:	good morning
English:	may i speak to the head of the household
Arabic:	i'm the owner of the family and i can speak with you
English:	may i speak to you about problems with your utilities
Arabic:	yes i have problems with the utilities

Table 1: Sample dialogue from the IraqComm Corpus.

tator contains exactly one ASR error. We use a crowdsourcing resource, Amazon Mechanical Turk (AMT) [7], to obtain human judgments about error recovery strategies for these utterances.

3.2. Method

The experiment is text-based. We gave each AMT worker an original user utterance from the dataset's manual transcript. The words misrecognized by the automatic speech recognizer were replaced by "XXX" to indicate a recognition error.[3] This is intended to simulate a dialogue system's automatic detection of misrecognized words in an utterance. We ask the AMT workers to answer a set of questions about their perception of the misrecognized utterance and then ask them how they would try to recover the missing information for the sentence. Table 2 shows a sample sentence and questions presented to the participants. Each sentence was presented to three AMT workers.

Original user utterances with an ASR error	
	how many XXX doors does this garage have
Questions to participants	
1.	Is the meaning of the sentence clear to you despite the missing word?
2.	What do you think the missing word could be? If you're not sure, you may leave this space blank.
3.	What type of information do you think was missing?
4.	If you heard this sentence in a conversation, would you continue with the conversation or would you stop the other person to ask what the missing word is?
5.	If you answered "stop to ask what the missing word is", what question would you ask?

Table 2: *Questions given annotators.*

From this annotation we are able to investigate human strategies for 1) the choice of action: continue dialogue or engage in clarification; 2) the type of clarification question (reprise vs. non-reprise), and 3) the grammatical structure of the reprise questions they produce. Below we discuss results from an initial analysis of this data.

[3] In the current dataset each error contains exactly one misrecognized word. We are now collecting data where multiple words may have been misrecognized.

3.3. Results

For each input sentence, the annotators had to decide first whether they would continue the conversation without interruption or ask a question about the missing information. If they chose to ask a question, they were prompted to construct an appropriate question. Table 3 shows examples of annotator decisions and clarification questions for several sample sentences. In Example 1, a noun at the end of the sentence is missing and two of the annotator choose to ask a reprise clarification question, while one annotator chooses to continue without clarification. In Example 2, a verb in the beginning of the sentence is missing and two of the annotators choose to continue while one chooses to ask a generic clarification question. In Example 5, one of the annotators asks a clarification question — erroneously assuming that the missing word is an adjective.

POS tag	**num/% in dataset**	**Correct POS**	**Correct word**
noun	101 (21%)	70%	10%
verb	133(28%)	50%	48%
pronoun	25 (5%)	73%	48%
adjective	34 (7%)	55%	22%
adverb	8 (2%)	29%	4%
preposition	34 (7%)	69%	51%
wh-question	48(10%)	75%	64%
other	92 (19%)	-	31%
overall		49%	39%

Table 4: Percentage of correctly hypothesized POS tags/words.

Annotatorss were also asked guess the identify of the missing word and its POS tag. When guessing POS tags, annotator were given a closed set of tags: name/place, noun, verb, pronoun, adjective, adverb, preposition, wh-question, other. They were also given examples for each tag. Table 4 shows the distribution of POS tags among misrecognized words as well as annotator accuracy in guessing correct word and tag. Overall accuracy for POS tag hypotheses in our dataset is 49% and accuracy of word identification is 39%. These results indicate that humans are indeed sometimes able to fill in missing content. This suggests that, to recover from a speech recognition error, a system should first attempt to hypothesize the misrecognized word before asking a clarification question. Our results show that, when a missing word is a verb or a closed-class word, such as a pronoun, a *wh*-word, or a preposition, a human is especially likely to guess correctly. In our data, they guess the POS of 73% of pronouns POS, but actual word identity only 48% of the time. Percentages of verb POS tags and verbs hypothesized very similar to one another (50% / 48%), indicating that most annotatorss who can guess that a missing word

id	Sentence	POS tag	Word	Annotator Decisions	Annotator Question(s)
1.	do you own a XXX	noun	hardhat	Continue(1), RepriseQ(2)	Do I own a what?/ Do I have what?
2.	XXX these actions successful	verb	were	Continue(2),GenericQ(1)	What did you say?
3.	make sure you close the XXX behind the vehicle	noun	door	RepriseQ(3)	Close the what?/What needs to be closed?/What behind the vehicle needs to be closed?
4.	how long have the villagers XXX on the farm for	verb	lived	Continue (3)	-
5.	XXX signs on the road are very important	verb	having	Continue(2), Reprise(1)	I'm sorry what type of signs?

Table 3: Sample annotator responses.

is a verb can also guess the word itself. In our dataset, most misrecognized verbs are auxiliary verbs *"to be", "to do", "to have"*, which may be easier to guess than other verbs. Nouns POS tags, on the other hand, were correctly guessed in 70% of cases but the nouns themselves were rarely (10% of cases) identified indicating, not surprisingly, that a clarification question for nouns is desirable in open domain systems.

POS Hyp.	Continue no Q%	Generic Q	Conf. Q	Repr. Q
name/place	23%	5%	5%	68%
noun	27%	11%	4%	58%
verb	62%	6%	2%	30%
pronoun	69%	3%	5%	23%
adjective	24%	4%	4%	45%
adverb	68%	5%	7%	20%
prep	85%	2%	4%	8%
wh-q	86%	5%	2%	6%
other	61%	13%	1%	25%
overall	60%	7%	3%	30%

Table 5: Annotator responses to missing data.

Table 5 shows the distribution of annotator responses to missing information in our dataset. Overall, in 60% of cases annotators choose to continue without a clarification question; in 30%, they ask a reprise clarification queston; in 7%, they ask a generic clarification (e.g. *"Please repeat."*; and in 3% of cases they ask a confirmation question (e.g. "Did you say...". The distribution of each decision type varies for different annotator hypotheses about a word's POS tag. Reprise clarification questions are asked in 58% of cases where an annotator guesses the POS tag to be a noun, but only in 6% of cases where a annotator guesses the POS tag to be a wh-word.

4. Conclusions and Future Work

In this study we have presented a preliminary analysis of a corpus of utterances containing ASR errors, annotated by Amazon Mechanical Turk works for POS and identity of the misrecognized word, as well as annotators' likely response to such errors: continue without clarification; generic request to repeat, rephrase, or confirm; or reprise clarification question. In over 60% of cases, annotators choose to continue the dialogue *without* asking clarification. For some categories of errors (auxiliary verbs and function words), annotators could hypothesize the missing words with good accuracy. This suggests that spoken dialogue systems might avoid sometimes risky clarification subdialogues by making use of syntactic information to also hypothesize misrecognized words. Similarly to previous studies, we found targeted reprise clarifications to be the most common kind of clarification question. However, we also found that humans are much more likely to propose a reprise clarification question when they believe the missing word to be a noun than another POS, suggesting that systems should focus their strategies for constructing such questions on that category.

In future work, we will use these annotations to train statistical models for identifying when a dialogue system should or should not engage in a clarification dialogue and what type of clarification question should be presented to a user. Features we think will be important in this modeling are POS as well as semantic and dependency parse information We will incorporate this classifier into an automatic clarification question generation tool to construct natural clarification questions. Our immediate application for this tool goal is to improve the clarification engine of a speech-to-speech translation system based.

5. References

[1] M. Akbacak et al. Recent advances in SRI's IraqComm^{tm} Iraqi Arabic-English speech-to-speech translation system. In *ICASSP*, pages 4809–4812, 2009.

[2] A. Rudnicky D. Bohus. Sorry, i didn't catch that! - an investigation of non-understanding errors and recovery strategies. In *Proceedings of the 6th SIGdial Workshop on Discourse and Dialog*, 2005.

[3] D. Klein and C. D. Manning. Accurate unlexicalized parsing. In *Proceedings of the 41st annual meeting annual meeting of the Association for Computational Linguistics*, pages 423–430, 2003.

[4] T. Koulouri and S. Lauria. Exploring miscommunication and collaborative behaviour in human-robot interaction. In *SIGDIAL Conference*, pages 111–119, 2009.

[5] M. Purver. *The Theory and Use of Clarification Requests in Dialogue*. PhD thesis, King's College, University of London, 2004.

[6] G. Skantze. Exploring human error recovery strategies: Implications for spoken dialogue systems. *Speech Communication*, 45(2-3):325–341, 2005.

[7] Amazon Mechanical Turk. http://aws.amazon.com/mturk/, accessed on 28 may, 2012.

[8] B. A. Weiss et al. Performance evaluation of speech translation systems. In *LREC*, 2008.

[9] J. D. Williams and S. Young. Characterizing task-oriented dialog using a simulated ASR channel. In *Proceedings of the ICSLP, Jeju, South Korea*, 2004.