Detecting Inappropriate Clarification Requests in Spoken Dialogue Systems

Alex Liu[1], Rose Sloan[2], Mei-Vern Then[1], Svetlana Stoyanchev[3], Julia Hirschberg[1], Elizabeth Shriberg[4]
Columbia University[1], Yale University[2], AT&T Labs Research[3], SRI International[4]
{al3037@columbia.edu, rose.sloan@yale.edu,
mt2837@columbia.edu, sveta@research.att.com,
julia@cs.columbia.edu, elizabeth.shriberg@sri.com}

Abstract

Spoken Dialogue Systems ask for clarification when they think they have misunderstood users. Such requests may differ depending on the information the system believes it needs to clarify. However, when the error type or location is misidentified, clarification requests appear confusing or *inappropriate*. We describe a classifier that identifies inappropriate requests, trained on features extracted from user responses in laboratory studies. This classifier achieves 88.5% accuracy and .885 F-measure in detecting such requests.

1 Introduction

When Spoken Dialogue Systems (SDS) believe they have not understood a user, they generate requests for clarification. For example, in the following exchange, the System believes it has misunderstood the word *Washington* in the user's utterance and asks a clarification question, prompting the user to repeat the misrecognized word.

User: I'd like a ticket to *Washington*.
System: A ticket to where?
User: Washington.

Clarification requests may be generic or specific to the type and location of the information the system believes it has not recognized. **Targeted clarifications** focus on a specific part of an utterance, as in the system's question above. They use understood portions of an utterance (*"I'd like a ticket to"*) to query a misunderstood portion (*"Washington"*). Targeted clarification is a type of task-related request, which has been shown to be more effective and prevalent in human-human dialogues than more general clarification requests (Skantze, 2005). Such **generic clarifications** signal misunderstanding without identifying the type or location of the misunderstanding. They often take the form of a request to repeat or rephrase, e.g. *"please repeat"*, *"please rephrase"*, *"what did you say?"*.

Questions that address a particular type of misrecognition come in several varieties. Systems may ask reprise clarification questions, by repeating a recognized portion of an utterance (Ginzburg and Cooper, 2004; Purver, 2004). Systems may also request that users spell a word if they believe the misrecognized word is a proper name, especially one that is not in its vocabulary (OOV). They may ask the user to provide a synonym for OOV terms that are not proper names. Systems may also ask users to disambiguate homophones (e.g. "Did you mean 'right' as in correct or 'rite' as in a ritual?"). They may request confirmation explicitly (e.g. "I heard you say Washington. Is that correct?"), or implicitly, by repeating the recognized information while asking a follow-up query (e.g. "When do you want to go to Washington?"). Each request type may be appropriate in different circumstances. However, when systems make *inappropriate* requests to users, such as to rephrase a proper name or to confirm a statement that contains a misrecognized word, dialogues often go awry. Therefore, it is extremely important for systems to know when a request is *inappropriate*, so that they can provide a different clarification request or fall back to a more generic strategy.

In this work, we develop a data-driven method for detecting inappropriate clarification requests. We have defined a list of inappropriate request types and have collected a corpus of speaker responses to both appropriate and inappropriate requests under laboratory conditions. We use this corpus to train an inappropriate clarification classifier to be used by a system after a user responds to a system request, in order to determine whether the question was appropriate or not. In Section 2, we describe previous research on error handling in dialogue. We describe our data set in Section 3 and

Proceedings of the SIGDIAL 2014 Conference, pages 238–242,
Philadelphia, U.S.A., 18-20 June 2014.

our approach in Section 4. We present our evaluation results in Section 5. We conclude in Section 6 and discuss future directions.

2 Related Work

Today's SDS use generic approaches to clarification, asking the user to repeat or rephrase an entire utterance when the system believes it has not been understood correctly. They use *confidence scores* on the ASR hypothesis to decide whether to accept, reject, or ask for clarification (Bohus and Rudnicky, 2005). Hypotheses with low scores may be confirmed and those with lower scores will trigger a generic request for repetition or rephrasing. Researchers have found that the formulation of system prompts has a significant effect on the success of SDS interaction. Goldberg et al. (2003) find that form of a clarification question affects user frustration and the consequent success of clarification subdialogue. In previous work, we explored the use of targeted reprise clarifications to improve naturalness (Stoyanchev et al., 2014).

Lendvai et al. (2002) apply machine learning methods to detect errors in human-machine dialogue, focusing on predicting when a user utterance causes a misunderstanding. Litman et al. (2006) identify user corrections of the system's recognition errors from speech prosody, ASR confidence scores, and the dialogue history. In contrast, we focus here on detecting when a **system** clarification request is the cause of dialogue problems. We employ only lexical features here, as well as the type of system request, to investigate user responses to a wide variety of system requests, and to identify system errors in request formulation from user reactions. In future work we will include acoustic and prosodic features as well.

3 Data

Our data consists of spoken answers to clarification requests collected at Columbia University using a simulated dialogue system in order to control recognition results and type of system response. The system displays a sentence and asks the user to read it. The system then issues a pre-prepared clarification request, which may be appropriate or inappropriate, to which the user responds. For example, in the following exchange, the system simulates a misunderstanding of the word *furor* by asking a targeted reprise clarification question.

User: We hope this won't create a *furor*.
System: Create a what?
User: A furor, an uproar.

The system issued six different types of clarification requests: confirmation; rephrase, spell, or disambiguate part of the utterance; targeted reprise clarification; and a targeted-reprise-rephrase combination. These request types were chosen based on the types of requests made by the SRI ThunderBOLT speech-to-speech translation system (Ayan and others, 2013). Confirmation questions simply ask the user to confirm an ASR hypothesis. Rephrase-part requests ask users to rephrase a specific part of an utterance which is played back to the user. Spell questions ask users to spell a word or phrase using the NATO alphabet. Disambiguate questions clarify ambiguous terms. Targeted reprise clarification questions make use of the recognized portion of an utterance to query the part that has been misrecognized based on the system's assessment. Targeted-reprise-rephrase requests are similar, with the additional request for the user to rephrase a portion of the utterance believed to have been misrecognized, which is played to the user.

Inappropriate requests in this study were defined as those that resulted from the ThunderBOLT system's incorrect identification of an error segment or an error type. For example, the clarification request *"Please say a different word for Afdhal"* is inappropriate since it asks for a rephrasal of a proper name. A request to spell a very long phrase is also identified as inappropriate since users have found this difficult, especially when using the NATO alphabet. Requests to disambiguate in the system provide two possible senses of the ambiguous word and are inappropriate when the correct sense is not one of the two provided. Targeted reprise clarification questions are inappropriate when the error segment is not correctly recognized and an errorful segment is included in the question (e.g. *"The okay I zoo would like what?"*). An appropriate question correctly identifies the error segment or ambiguous term and the error type. For example, the question *"I think 'Afdhal' is a name. Please spell it"*, would be appropriate when 'Afdhal' is OOV because it correctly targets the error and its type.

For each clarification request type, except for confirmation questions, which are always appropriate, we created one or more types of *inappropriate* requests for each of the conditions we ob-

served in dialogues collected with the Thunder-BOLT system. For example, when the system asks the user to rephrase a part of their utterance which the system believes to be a misrecognized non-proper-name, the question is appropriate when indeed that non-proper-name has been misrecognized. However, the request will be inappropriate when the hypothesized error segment played back to the user is a partial word, a proper name, an extended segment including a name, or a function word. We created instances of each of these conditions for our users to respond to in our experiment. A full list of the system question types and their appropriate and inappropriate conditions is provided in Table 3, in the Appendix. We prepared 228 clarification requests (84 appropriate and 144 inappropriate), 12 for each of the 19 categories listed in Table 3 in the Appendix, based on data in the TRANSTAC dataset (Akbacak and others, 2009). Our subjects were 17 native American English speakers, each of whom answered 114 requests. We recorded speakers' answers to 714 appropriate and 1224 inappropriate requests. As most request types have more than one inappropriate version, 63% of the requests in the data set are inappropriate.

4 Experiment

We used the Weka machine learning library (Witten and Eibe, 2005) to train classifiers to predict whether a clarification request was appropriate or inappropriate. Our features were extracted from transcripts of user utterances, and included lexical, syntactic, numeric, and features from the output of Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2007) as described in Table 1.

We included unigram and bigram features, excluding unigrams that appeared fewer than 3 times in the dataset (11% of the unigrams), and bigrams that appeared fewer than 2 times (25%), with thresholds set empirically. LIWC features were extracted using the LIWC 2007 software, which includes lexical categories, such as articles and negations, and psychological constructs, such as affect and cognition. In one version of the corpus, we replaced sequences of user spellings with the tag "SPELL" and disfluencies with the symbol "DISF". We used the Stanford POS tagger (Toutanova and others, 2003) to tag both the original corpus as well as the modified version. In the latter, we replaced the "SPELL" and "DISF" tags with the symbols themselves. We also mapped nine of the most frequent unigrams to their own POS classes, such as "no", "not", and "neither" to "NO" and "word" to "WORD". We then used counts of POS bigrams as a syntactic feature. Additionally, as we observed that responses to inappropriate requests contained a higher proportion of function words, we added this as a numeric feature. We also observed that average length of responses to inappropriate requests was greater than responses to appropriate ones, and we hypothesized this was in part due to inappropriate requests to spell long phrases. Therefore, we also used the length of the total spelling sequences, or the count of letters spelled out, as a numeric feature. We also added type of clarification request as a feature since some requests are less likely to be inappropriate than others. For example, we consider confirmation questions ("Did you say . . . ?") to always be appropriate.

Feature	Description
word_unigrams (Lexical)	Count of unigrams
word_bigrams (Lexical)	Count of bigrams
pos_bigrams (Syntactic)	Bigrams of POS assigned by Stanford tagger
liwc	LIWC Output
func_ratio	Proportion of function words in response
len_spell	Total length of spelling sequences in response
request_type	Type of request preceding response

Table 1: *Features used in Classification.*

5 Results

We report classification results using Weka's J48 decision tree classifier with 10-fold cross validation in Table 2, which outperformed JRip and LibSVM in our experiments. Compared to the majority baseline of 63.2% accuracy and .489 F-measure, our classifier which uses all of the features in Table 1 achieves a significant improvement, with an accuracy of 88.5% and an F-measure of .885. A baseline method that uses only system request type feature (Req. type baseline) achieves accuracy of 73.7% and F-measure of .686, which is significantly below the performance of the trained classifier. To identify the most important features in predicting inappropriate requests, we iteratively removed a single feature from the full feature set and re-evaluated prediction accuracy. Table 2 shows absolute decrease

Features	Acc (%)	P/R/F-Measure
Majority baseline	63.2 *	0.399/0.632/0.489
Req. type baseline	73.7 *	0.814/0.737/0.686
All Features	88.5	0.885/0.885/0.885
less request_type	−7.6 *	−0.076
less liwc	−2.3	−0.023
less pos_bigrams	−2.0	−0.020
less word_unigrams	−0.4	−0.004
less func_ratio	−0.1	−0.001
less len_spell	−0.05	−0.0005
less word_bigrams	+0.05	+0.0007

Table 2: *Classifying Inappropriate Requests: All Features vs. Baseline vs. Leave-One-Out Classifiers, where * indicates statistically significant difference from All Features ($p < 0.01$)*

in percentage points and in F-measure when each feature is removed in turn compared to the classifier trained on the full features set. We found that system request type was the most important feature, as performance decreased by 7.6 percentage points without it. This makes sense in light of the fact that the ratio of inappropriate to appropriate requests varied for the different request types represented in our dataset. The next most useful features were the output of LIWC and the POS bigrams. We had hypothesized that, since LIWC captures the presence of negations and assents, it could capture negative user responses to the system such as *yes* or *no*. As for the POS bigrams, we modified the POS tags to mark common words and included start and end markers in the bigrams because we hypothesized that the first words and last words in the responses might be particularly informative. Looking at the decision tree created with all our features, we find that the first five branches involve decisions regarding the unigrams "name" and "SPELL" (a collapsed spelling sequence), the ⟨START, "neither"⟩ bigram, the LIWC ingestion-word feature, and the type of request, in that order. Not only do these findings confirm our hypotheses, they also confirm that the unigrams "name", "SPELL", and "neither" which we had mapped to special POS classes are particularly useful.

After training our model, we used it to classify our entire dataset to see which responses it performed well on and which it tended to misclassify. Responses to targeted reprise and targeted-reprise-rephrase questions together accounted for around half of the misclassified instances. Many easily identifiable responses to inappropriate requests involved the user correcting the system, as in the following example:

User: You are going to need to dole out punishment.
System: I think this is a name: 'dole out punishment'. Please spell that name.
User: It is not a name, it is a phrase, dole out punishment.

However, when the users did **not** correct the system after an inappropriate request, their responses appeared no different from answers to appropriate requests. In the following example, the system misrecognizes "hyperbaric" and interprets it as the word "hyper" followed by an unknown phrase, but the user simply ignores the request and repeats.

User: We are going to put you in a hyperbaric chamber.
System: Put you in a high what? Please give me another word or phrase for 'perbaric'.
User: Hyperbaric chamber.

Many cases in which appropriate requests were misclassified as inappropriate involved users responding correctly to targeted or targeted-rephrase questions. We hypothesize that these are also due primarily to users ignoring the inappropriate system request and providing the information the system **should have** asked for. As a result, those cases make it difficult to distinguish between responses to appropriate and inappropriate targeted questions. Of course, users may be giving prosodic cues to indicate confusion or uncertainty or hyperarticulating in their responses. We will address the use of prosodic features in predicting inappropriate requests in future work.

6 Conclusions

In this work, we have addressed a novel task of identifying inappropriate clarification requests using features extracted from user responses. We collected responses to inappropriate clarification requests based on six request types in a simulated SDS environment. The classifier trained on this dataset detects inappropriate requests with accuracy of 88.5%, which is 25.3 percentage points above the majority baseline, and an F-measure of .885, which is .396 points above the majority F-measure. In future work, we will include acoustic and prosodic features as well as lexical features and we will evaluate the use of an inappropriate clarification request component in our speech-to-speech translation system.

References

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

N. F. Ayan et al. 2013. "Can you give me another word for hyperbaric?": Improving speech translation using targeted clarification questions. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 8391–8395. IEEE.

D. Bohus and A. I. Rudnicky. 2005. A principled approach for rejection threshold optimization in spoken dialog systems. In *INTERSPEECH*, pages 2781–2784.

J. Ginzburg and R. Cooper. 2004. Clarification, ellipsism and the nature of contextual updates. *Linguistics and Philosophy*, 27(3).

J. Goldberg, M. Ostendorf, and K. Kirchhoff. 2003. The impact of response wording in error correction subdialogs. In *ISCA Tutorial and Research Workshop on Error Handling in Spoken Dialogue Systems*.

P. Lendvai, A. van den Bosch, E. Krahmer, and M. Swerts. 2002. Improving machine-learned detection of miscommunications in human-machine dialogues through informed data splitting. In *Proceedings of the ESSLLI Workshop on Machine Learning Approaches in Computational Linguistics*, pages 1–15.

D. Litman, J. Hirschberg, and M. Swerts. 2006. Characterizing and predicting corrections in spoken dialogue systems. *Computational linguistics*, 32(3):417–438.

J. W. Pennebaker, C. K. Chung, M. Ireland, A. Gonzales, and R. J. Booth, 2007. *The development and psychometric properties of LIWC2007*. Austin, TX.

M. Purver. 2004. *The Theory and Use of Clarification Requests in Dialogue*. Ph.D. thesis, King's College, University of London.

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

S. Stoyanchev, A. Liu, and J. Hirschberg. 2014. Towards natural clarification questions in dialogue systems. In *Proceedings of AISB2014*.

K. Toutanova et al. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*. Association for Computational Linguistics.

I. Witten and F. Eibe. 2005. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, 2nd edition.

Appendix

ID	Simulation	Appro.	Example
1. Confirmation			
1	Correctly recognized utterance	yes	Did you say "place this on the pane"?
2	Misrecognized utterance	yes	Did you say "these are in um searches will cause the insurgents to priest buyer"?
2. Rephrase-part			
1	Full non-name word or phrase	yes	Please say a different word for "surmise".
2	Partial word	no	Please say a different word for "nouncing".
3	Name	no	Please say a different word for "Afdhal".
4	Extended segment including name	no	Please say a different word for "checkpoint at Betirma".
5	Function word	no	Please say a different word for "off over".
3. Disambiguate			
1	One choice is correct	yes	Did you mean fliers as in handouts or fliers as in pilots?
2	Neither choice is correct	no	Did you mean plane as in aircraft or plain as in simple?
3	Word being disambiguated was not said	no	Did you mean sight as in vision or site as in location?
4. Spell			
1	Name	yes	Please spell "Hadi Al Hemdani".
2	Non-name	no	I think this is a name: "eluding". Please spell that name.
3	Extended segment	no	Please spell "staff are stealing themselves".
5. Reprise			
1	Error segment correctly recognized and no other errors	yes	We will search some of the what?
2	Recognition error right before "what" word	no	Supplies of I see them what?
3	Recognition error which is not the last word before "what"	no	Ask if they are for eating for what?
6. Reprise rephrase			
1	No errors outside of the error segment	yes	Use a what? Please say another word for "bristled".
2	Error segment is a partial word	no	Are there any my what? Please say another word for "nors".
3	Error outside the targeted segment	no	Be a right is what? Please say another word for "rain".

Table 3: *Clarification Requests and Contexts in which they are Appropriate and Inappropriate.*