# Exploring Features For Localized Detection of Speech Recognition Errors

**Eli Pincus** and **Svetlana Stoyanchev** and **Julia Hirschberg**
Department of Computer Science, Columbia University, USA
elipincus@gmail.com & sstoyanchev@cs.columbia.edu
& julia@cs.columbia.edu

## Abstract

We address the problem of *localized error detection* in Automatic Speech Recognition (ASR) output to support the generation of *targeted clarifications* in *spoken dialogue systems*. Localized error detection finds specific mis-recognized words in a user utterance. Targeted clarifications, in contrast with generic 'please repeat/rephrase' clarifications, target a specific mis-recognized word in an utterance (Stoyanchev et al., 2012a) and require accurate detection of such words. We extend and modify work presented in (Stoyanchev et al., 2012b) by experimenting with a new set of features for predicting the likelihood of a local error in an ASR hypothesis on an unsifted version of the original dataset. We improve over baseline results, where only ASR-generated features are used, by constructing optimal feature sets for utterance and word mis-recognition prediction. The f-measure for identifying incorrect utterances improves by 2.2% and by 3.9% for identifiying incorrect words.

## 1 Introduction

Spoken Dialogue Systems typically indicate their lack of understanding of user input by simple requests for repetition or rephrasing – "I'm sorry, I didn't understand you.", or "Can you please repeat?". However human conversational partners generally provide more targeted clarification requests. Corpus analysis of human conversations have shown that people are more likely to indicate what they have understood and what they have *not* understood by producing *reprise clarification questions* (Purver, 2004; Stoyanchev et al., 2012a), as illustrated in the following exchange where XXX indicates a word misunderstood by speaker B:

A: Do you have any XXX in your bag?
B: Do I have any what in my bag?

A reprise clarification question targets a specific mis-recognized word and incorporates recognized context into a clarification question.

We investigate replacing generic *please repeat* clarifications with more natural targeted clarifications in automatic spoken systems. Targeted clarifications allow users to provide a concise response to a clarification question which is beneficial for spoken systems accepting broad vocabulary and flexible syntax. Examples of such systems include tutoring systems, intelligent assistants, and spoken translation systems (Litman and Silliman, 2004; Dzikovska et al., 2009; Akbacak et al., 2009).

To enable Spoken Dialogue Systems (SDS) to generate targeted clarification questions, we must first be able to identify mis-recognized words with high accuracy. We term such mis-recognition detection *localized error detection*. Accurate distinction between correctly and incorrectly recognized words is essential to the creation of appropriate targeted clarification questions.

In previous research on recognition error detection in dialogue systems, researchers have addressed error detection at the utterance level (Hirschberg et al., 2004; Komatani and Okuno, 2010). In this paper we present results of classification experiments designed to detect localized errors within the utterance. Our baseline results are obtained from a classifier trained only on word posterior probabilities generated by an Automatic Speech Recognition (ASR) engine. ASR confidence score computation is an active research area, relying upon acoustic and lexical collocation information to compute confidence scores. We determine whether improvement over baseline can be achieved by training a classifier for utterance and word mis-recognition prediction on an expanded feature set that includes lexical, positional, prosodic, semantic, syntactic as well as additional ASR score features. All of the features we experiment with can be computed from an ASR hypothesis without affecting the performance of a SDS materially. After determining optimal feature sets we experiment with one- and two-stage approaches for localized error detection. The first simply identifies whether a word is correctly recognized or not. The second first classifies an utterance as incorrect or correct and then classifies errors only on utterances labeled incorrect.



*Proceedings of the SIGDIAL 2013 Conference*, pages 132–136,
Metz, France, 22-24 August 2013. 

This work extends earlier work in which we evaluated a smaller set of syntactic and prosodic features (Stoyanchev et al., 2012b). In addition to improvements implemented in the ASR engine that we use to produce ASR hypotheses, our current work reports results on a larger dataset which includes commands to the system and utterances containing disfluencies. Here, we propose a framework for localized error detection that does not rely upon pre-filtering of the dataset.

In Section 2 we describe our corpus. In Section 3 we discuss our classification experiments. In Section 4 we discuss our results. In Section 5 we present our conclusions and discuss future research.

## 2 Data

We conduct our machine learning experiments on the DARPA TRANSTAC corpus (Weiss et al., 2008). The TRANSTAC corpus is comprised of staged conversations between American military personnel and Arabic interviewees utilizing IraqComm speech-to-speech translation system (Akbacak et al., 2009). This data was collected by NIST between 2005 and 2008 in evaluation exercises. The dataset contains audio recordings and manual transcript of English and Arabic utterances. We used SRI's DynaSpeak (Franco et al., 2002) speech recognition system to recognize the English utterances and use posterior probabilities from DynaSpeak as our baseline feature. We create a corpus from this dataset that contains over 99% of the English utterances. 38 utterances were removed from the dataset either for lack of actual speech data or errors in reference transcription. 26.2% of our cleaned corpus consist of mis-recognized instances and 6.4% of the total words in it are incorrectly recognized by DynaSpeak (see Table 1). We are using an unsifted version of the corpus used in our previous work (Stoyanchev et al., 2012b) whose hypotheses were produced with a new version of the DynaSpeak ASR system. In our previous work utterances containing disfluencies and commands to the system were excluded. We seek to avoid the cascading errors that would follow from implementing a 2-step framework for localized error detection where the first step is command and disfluency detection and the second step is localized error detection. The 1-step framework also has the advantage of working for all utterances including ones that contain commands or disfluencies. Due to these differences, our current results are not directly comparable with our previous results.

Table 1: *Corpus statistics*

| | **Overall** | **Correct ASR** | **Incorr ASR** |
|---|---|---|---|
| All utts. | 3,952 | 2,914 (73.7%) | 1,038 (26.2%) |
| All wrds. | 25,333 | 23,705 (93.6%) | 1,628 (6.4%) |
| wrds in err utts | 7,888 | 6,260 (79.4%) | 1,628 (20.6%) |

## 3 Method

We analyze how the performance of predicting mis-recognized utterances and words is affected by the use of lexical, positional, prosodic, semantic, and syntactic features in addition to ASR confidence scores. We perform machine learning experiments using the Weka Machine Learning Library to construct a J48 decision tree classifier boosted with MultiBoostAB method (Witten and Eibe, 2005).

**Baseline confidence features** We use ASR posterior scores extracted from the log files output by Dynaspeak as a baseline feature set in our experiments. In the utterance mis-recognition prediction experiment, we calculate the average of the logarithm of the ASR posterior scores over all words in the hypothesis. In the word mis-recognition prediction experiment we use the logarithm of the posterior score of a given word.

**Feature selection** We run a heuristic feature exploration experiment to identify optimal feature sets for predicting mis-recognized utterances and mis-recognized words. We first use a greedy approach adding one feature at a time to the baseline ASR feature set and only keep a feature in the set if it improves F-measure predicting mis-recognition. We then use an alternate greedy approach in which we begin with a feature set composed of all extracted features and proceed to remove one feature at a time and only leave it out of the set if incorrect F-measure improved or remained the same with its absence. The second approach yields the optimal feature sets for both utterance and word mis-recognition prediction. Table 2 lists the features that make up these optimal sets. For incorrect utterance prediction, we run a 10-fold cross validation on all utterances. For incorrect word prediction, we run a 10-fold cross validation on all words in mis-recognized utterances.[1] We next describe the features we found to be useful in prediction and those that did not improve performance.

### 3.1 Useful Features

**ASR context features** We use the logarithm of the posterior score of a given word and the average of the logarithm of the posterior scores for both a given word and its surrounding context. We use one word context before and after the given word. We also use the average of the logarithm of the posterior scores for all words in the utterance.

**Lexical features** We hypothesize that properties of words such as length and frequency are predictive of whether a word is correctly recognized. In particular, noting that words of greater length are often better recognized by an ASR engine, we examine the length, frequency, and posterior score of the maximum and min-

[1] Because of the size limitations of our dataset feature selection and evaluation are performed on the same dataset.

imum words in an utterance. For mis-recognized utterance prediction, we find that the average length of a word in the utterance are useful features for predicting both mis-recognized utterances and words. For mis-recognized word prediction, we find the word length of the surrounding words, the current word, and the frequency of the longest word in an utterance are useful. We also find that *utterance length* calculated in words is a useful feature for predicting both utterance and word mis-recognition.

**Positional features** Motivated by the use of dialogue history features in Lopes et al (2011), we find that the location of the hypothesis relative to the speaker's first utterance in the dialogue (*utterance location*) is a useful feature. Similarly, we obtain improvement from the *word index* feature, the distance of the word from the first word in the utterance.

**Syntactic** POS tags were shown to be helpful in our previous work and we find that these tags improve the current results as well. We obtain these from the Stanford POS tagger (et al., 2003). In mis-recognized utterance prediction, we use unigram and bigram counts of POS tags as a feature. For mis-recognized word prediction, we use the word's POS tag as well as the POS tag for the surrounding one or two words.

We obtain a binary *Func/Content* feature using a function word list to distinguish function from content words. The list includes certain adverbs, conjunctions, determiners, modal verbs, primary verbs such as *be*, prepositions, pronouns, and WP-pronouns. These tags also boost our ability to identify mis-recognized words. The feature *Func/Tot ratio* is the fraction of function words to total words in an ASR hypothesis. We hypothesize that an extreme value of the *Func/Tot ratio* may indicate a potential mis-recognition, and it does improve both utterance and word mis-recognition prediction.

### 3.2 Less Useful Features

Features we do not find helpful include information associated with the minimum length word in the utterance, the fraction of words in an utterance that possess greater length than the average length word in the corpus, as well as syntactic features such as a dependency tag assigned to the word. Additional unhelpful features include prosodic features, such as shimmer and jitter identified by PRAAT (Boersma and Weenink, 2013) and pitch and phrase information extracted from AuToBI(Rosenberg, 2010) software. Performing a semantic role label of our hypotheses with the software SENNA (Collobert et al., 2011) also did not provide helpful semantic features.

**System Performance** To evaluate performance of our mis-recognized word classifier, we use the selected features in 1-stage and 2-stage approaches. First, we train models for utterance and word classification separately on 80% of the dataset with up-sampling (35%)[2] of the incorrect instances as well as with the actual distribution of incorrect instances in the corpus (20.6% utterances, 6.4% words). We then test these models on the remaining 20% of the dataset using the 1-stage and 2-stage approach. In the 1-stage approach we test on 20% of the total words in the corpus. In the 2-stage approach we first test on 20% of the total utterances in the corpus and then only test on the words in the utterances labeled as mis-recognized.

Table 2: *Features*

| **Cat** | **Specific** | **In Optimal Utt Feature Set** | **In Optimal Wrd Feature Set** |
|---|---|---|---|
| ASR | Log Post Score | Yes (avg of all wrds in utt) | Yes (curr wrd) |
| ASR-CTX | Log Post Score | No | Yes (avg of curr wrd, curr wrd context, avg of all wrds in utt) |
| Lex | Wrd length | Yes (avg wrd length in utt) | Yes (curr,prev,next) |
| | Max Wrd freq | No | Yes |
| | Utt length | Yes | Yes |
| POS | Utt location | Yes | Yes |
| | Word Index | No | Yes (curr) |
| Syn | POS Tag | Yes (unigram and bigram count) | Yes (curr,prev,next) |
| | Func/Cont tag | No | Yes (curr, prev, next) |
| | Func/Tot ratio | Yes | Yes |

## 4 Results

**New Feature Experiments** Using our newly constructed utterance feature set we are able to boost incorrect utterance classification F-measure by 2.2% from .597 to .610 (see Table 3). The increase in F-measure for incorrect utterance mis-recognition is due to an increase in incorrect utterance recall from .531 to .555. There is a slight decrease in incorrect utterance precision from .682 to .678. Overall classification accuracy improves by 2.1% points (absolute) from 81.2% to 83.3%. Using our newly constructed word feature set we are able to improve incorrect word classification F-measure by 3.9% from .620 to .644 (see Table 4). For incorrect word classification there is an increase in both mis-recognized word precision and recall; the former increasing from .678 to .719 and the latter increasing from .571 to .584. The results for incorrect word classification represent a statistically significant improve-

[2]This percentage was derived empirically.

Table 3: *Utterance new feature experiment results*

| Feature | Correct P — R — F | Incorrect P — R — F | % F-Measure Incorr Imp over ASR Only | Accuracy |
|---|---|---|---|---|
| ASR | .845 — .912 — .877 | .682 — .531 — .597 | - | 81.2% |
| ASR+LEX+POS+SYN | .851 — .906 — .878 | .678 — .555 — .610 | 2.2% | 83.3% |

Table 4: *Word new feature experiment results*

| Feature | Correct P — R — F | Incorrect P — R — F | % F-Measure Incorr Imp over ASR only | Accuracy |
|---|---|---|---|---|
| ASR | .893 — .930 — .911 | .678 — .571 — .620 | - | 85.5% |
| ASR+LEX+POS+SYN | .897 — .941 — .918 | .719 — .584 — .644 | 3.9% | 86.7% |

Table 5: *1-stage and 2-stage approach results*

| Experiment | Correct P — R — F | Incorrect P — R — F | Accuracy |
|---|---|---|---|
| Maj. Baseline | .94 — 1.00 — .97 | - — 0 — - | 94% |
| 1-stage original | .97 — .94 — .96 | .39 — .57 — .46 | 92% |
| 1-stage (35% upsample) | .98 — .90 — .94 | .31 — .72 — .44 | 89% |
| 2-stage original | .96 — .98 — .97 | .51 — .34 — .41 | 94% |
| 2-stage (35% upsample) | .96 — .96 — .96 | .41 — .46 — .43 | 93% |

ment[3]. Overall classification accuracy improves by 1.2% points (absolute) from 85.5% to 86.7%.

**1-stage and 2-stage experiments** To estimate how well a dialogue system could perform incorrect word classification we run our 1-stage and 2-stage approaches. The 1-stage approaches (with and without up-sampling) are able to achieve higher recall; while the 2-stage approaches (with and without up-sampling) are able to achieve higher precision. The 2-stage result's higher precision is not surprising given that this approach has two chances to filter out correct words — first with utterance classification and then with word classification. In our 1-stage approach with up-sampling we are able to identify almost 3/4 (72%) of the incorrect words in the corpus (see Table 5). In our 2-stage approach without up-sampling we are able to accurately label just over 1/2 (51%) of the total instances we identify as incorrect. In future work we will experiment with additional features in order to boost precision for incorrect word classification to a level suitable for use in the construction of reprise clarification questions.

## 5 Conclusions

We have presented results of machine learning experiments that utilize new features to improve localized detection of ASR errors to assist spoken dialogue system's production of reprise clarification questions. We conducted feature selection experiments to find optimal feature sets to train classifiers for utterance and word mis-recognition prediction. We find that certain lexical, positional, and syntactic features improve classification results over a baseline feature set containing only ASR posterior score features. We improve incorrect F-measure for utterance mis-recognition prediction by 2.2% by adding utterance length, location, fraction of function words to total words, average word length, and unigram and bigram count to the baseline feature set. By removing average word length as well as unigram and bigram count from this optimal set for utterances and adding the current word's ASR-context features, length, distance from first word, POS tag, Content/Function tag as well as the length of the current's words surrounding 1 or 2 word contexts, we improve incorrect F-measure for word mis-recognition prediction by 3.9% . We then employ these feature sets in 1-stage and 2-stage approached to obtain our final results. The 2-stage (no up-sampling) approach yields the highest precision for detection of word mis-recognition at 51% while the 1-stage (with 35% up-sampling) approach yields the highest recall for detection of word mis-recognition at 72%.

In order to implement this approach in a working dialog system we would need to increase our word mis-recognition precision. The presence of false positives in mis-recognition prediction (correctly recognized words classified as mis-recognized) could lead to unnecessary clarification requests — potentially derailing the dialogue.

In future work we will experiment with additional corpora as well as with an even more fine-grained approach to local error detection, looking for deletions, insertions, and substitutions. Potentially, optimal classifiers could be found for each of these types of mis-recognition. If we are able to identify the type of ASR error as well as its location, we should be able to improve our construction of clarifications questions.

We will also continue our investigation of how to use reprise clarification questions in SDS. Once we have detected localized ASR errors we must still refine our strategies for constructing clarification questions using this information. We are also studying how appropriate and inappropriate reprise clarification questions are handled by SDS users.

[3] $\chi^2 test(p < .01)$

## References


M. Akbacak, H. Franco, M. Frandsen, S. Hasan, H. Jameel, A. Kathol, S. Khadivi, X. Lei, A. Mandal, S. Mansour, K. Precoda, C. Richey, D. Vergyri, W. Wang, M. Yang, and J. Zheng. 2009. Recent advances in sri's iraqcomm; iraqi arabic-english speech-to-speech translation system. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 4809–4812.

P. Boersma and D. Weenink. 2013. Praat: doing phonetics by computer [computer program]. http://www.fon.hum.uva.nl/praat/.

R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. 2011. Natural language processing (almost) from scratch.

M. O. Dzikovska, C. B. Callaway, E. Farrow, J. D. Moore, N. Steinhauser, and G. Campbell. 2009. Dealing with interpretation errors in tutorial dialogue. In *Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, SIGDIAL '09, pages 38–45, Stroudsburg, PA, USA. Association for Computational Linguistics.

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*.

H. Franco, J. Zheng, J. Butzberger, F. Cesari, M. Fr, J. Arnold, V. Ramana, A. Stolcke R. Gadde, and V. Abrash. 2002. Dynaspeak: Sri's scalable speech recognizer for embedded and mobile systems. In *Proceedings of the second international conference on Human Language Technology Research*, HLT '02, pages 25–30, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

J. Hirschberg, D. J. Litman, and M. Swerts. 2004. Prosodic and other cues to speech recognition failures. *Speech Communication*, 43(1-2):155–175.

K. Komatani and H. G. Okuno. 2010. Online error detection of barge-in utterances by using individual users utterance histories in spoken dialogue system.

D. J. Litman and S. Silliman. 2004. Itspoke: an intelligent tutoring spoken dialogue system. In *Demonstration Papers at HLT-NAACL 2004*, HLT-NAACL–Demonstrations '04, pages 5–8, Stroudsburg, PA, USA.

J. Lopes, M. Eskenazi, and I. Trancoso. 2011. Towards choosing better primes for spoken dialog systems. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Keystone, CO.

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

A. Rosenberg. 2010. Autobi - a tool for automatic tobi annotation. In Takao Kobayashi, Keikichi Hirose, and Satoshi Nakamura, editors, *INTERSPEECH*, pages 146–149. ISCA.

S. Stoyanchev, A. Liu, and J. Hirschberg. 2012a. Clarification questions with feedback 2012. In *Interdisciplinary Workshop on Feedback Behaviors in Dialog*.

S. Stoyanchev, P. Salletmayr, J. Yang, and J. Hirschberg. 2012b. Localized detection of speech recognition errors. In *Spoken Language Technology Workshop (SLT), 2012 IEEE*, pages 25–30.

B. Weiss, C. Schlenoff, G. Sanders, M. Steves, S. Condon, J. Phillips, and D. Parvaz. 2008. Performance evaluation of speech translation systems. In Nicoletta Calzolari (Conference Chair) et al., editor, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

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