Context-based Speech Recognition Error Detection and Correction

Arup Sarma and David D. Palmer
Virage Advanced Technology Group
300 Unicorn Park
Woburn, MA 01801
dpalmer@virage.com

Abstract

In this paper we present preliminary results of a novel unsupervised approach for high-precision detection and correction of errors in the output of automatic speech recognition systems. We model the likely contexts of all words in an ASR system vocabulary by performing a lexical co-occurrence analysis using a large corpus of output from the speech system. We then identify regions in the data that contain likely contexts for a given query word. Finally, we detect words or sequences of words in the contextual regions that are unlikely to appear in the context and that are phonetically similar to the query word. Initial experiments indicate that this technique can produce high-precision targeted detection and correction of misrecognized query words.

1 Introduction

Spoken language sources, such as news broadcasts, meetings, and telephone conversations, are becoming a very common data source for user-centered tasks such as information retrieval, question answering, and summarization. Automatic speech recognition (ASR) systems, which can rapidly produce a transcript of spoken audio, are consequently becoming an essential part of the information flow. However, ASR systems often generate transcripts with many word errors, which can adversely affect the performance of systems designed to assist users in managing large quantities of natural language data. Retrieving documents or passages relevant to a user query is significantly easier when the words in the query are contained in the document; when a query word is misrecognized by the ASR system, retrieval accuracy declines. For example, if a user is searching for spoken documents related to "Iraq," and the spoken word "Iraq" is consistently misrecognized, the user will not be able to locate many of the desired documents.

In this work we introduce a novel unsupervised approach to detecting and correcting misrecognized query words in a document collection. Our approach takes advantage of two important patterns in the appearance of ASR errors. First, specific words in a large corpus tend to co-occur frequently with certain other context words, and misrecognitions of those specific words will also tend to co-occur with the same context words. Second, many ASR errors are phonetically similar to the actual spoken words. Our approach takes advantage of these patterns of ASR errors and seeks to find output words that are both phonetically similar to a query word and that occur in a context that is more likely to indicate the query word. For example, "Iraq" and "a rock" are phonetically very similar but generally occur in different contexts.

Our ASR error detection and correction is carried out in three steps that are separate from the speech recognition itself. We first analyze a large corpus of output from a given ASR system to compile co-occurrence statistics for each word in the system's vocabulary. This analysis results in a set of context words likely to occur with each vocabulary word. Next, given a target word, such as a query word entered into an information retrieval system, we identify regions in the search corpus containing a large number of the expected context words for the query word. Finally, we detect words in the regions that are unlikely to occur with the context words and that are phonetically similar to the query.

2 Our Approach

There are several key components to our approach to detecting and correcting in-vocabulary speech recognition errors. First, we calculate co-occurrence statistics for all words in a large corpus of ASR output data; this is an offline processing step that we describe in Section 2.1. This

co-occurrence information is used in an online error detection process based on word context analysis. The error detection process first requires the input of a query word that is to be sought in the test corpus of ASR output from the same engine; the goal of this step is to detect places in the corpus where the query word was spoken but misrecognized. We describe the contextual analysis in Section 2.2. From the set of candidate error regions, ASR errors are detected using phonetic comparison between the query word and words in the window; this phonetic analysis is described in Section 2.3.

Our approach to ASR error detection and correction builds on recent work in statistical lexical and contextual modeling using co-occurrence analysis, such as (Roark and Charniak, 1998). We apply the contextual modeling to a speech retrieval task, as in (Kupiec *et al.,* 1994). In the earlier work, general mathematical models were developed to measure lexical similarity between words in context. We seek to develop a simple contextual model based on word co-occurrences in order to facilitate the retrieval of spoken documents containing critical word errors.

Our approach has a similar goal to that of Logan (2002); however, their work focuses primarily on out-of-vocabulary words while we focus on in-vocabulary words. Our work also builds on recent directions in language modeling for speech recognition, in which a broader context beyond n-grams is considered. For example, the dimensionality reduction modeling of Bellegarda (1998) seeks to model long-range contextual similarity among words in a training corpus. Rosenfeld (2000) has developed another language modeling approach that can model word occurrences beyond the common trigram approaches. While language modeling techniques seek to improve the ASR engine itself, we present an ASR post-processing correction model, in which we process and improve the output of an ASR system.

The data used for our experiments consisted of a large corpus of English broadcast news transcripts produced by the broadcast news speech system described in (Makhoul *et al.,* 2000). This real-time ASR system has a vocabulary size of about 64k words and a reported performance that normally ranges from WER=20% to 30% for English news broadcasts. Our training corpus consisted of 360 half-hour broadcast transcripts containing roughly 1.6 million words. The broadcasts were from three different English sources (CNN Headline News, BBC America, and News World International) from July 2003. We divided the data into a training set, from which all model parameters were trained, and a separate test set consisting of files that were randomly selected from the corpus. The evaluation corpus consisted of 39 half-hour broadcast transcripts containing about 180,000 words.

2.1 Word Co-occurrence Analysis

The goal of the first step in our approach, co-occurrence analysis, is to determine, for any given word in the ASR vocabulary, the other words that are very likely to occur near the given word and are not likely to occur elsewhere. We compile co-occurrence frequencies for a target word by counting all other words that co-occur in a document with the target word within a certain window size *w* ($w/2$ words to the left and $w/2$ words to the right). In our case, we investigated windows sizes ranging from *w*=2 to *w*=40; as with all our system parameters, optimal value for a given source empirically through training.

We calculate several maximum likelihood prior probabilities for use in the co-occurrence analysis. For a target word x and each context word y, we calculate $p(x) = c(x)/n$ and $p(y) = c(y)/n$, where $c(x)$ and $c(y)$ are the total corpus counts for x and y and where n is the total number of words in the training corpus (1,638,224). We also calculate the joint probability $p(x, y) = c(x, y)/n$, the probability of co-occurrence for x and y in the training data. In addition, we calculate the pointwise mutual information $I(x, y)$ for two words x and y, $I(x, y) = \log \frac{p(x,y)}{p(x)p(y)}$. The value $I(x, y)$ is highest for target words x and context words y that occur frequently together within a window w but rarely outside the window. The context words are ranked by mutual information, and this ranked list of co-occurring context words for each target word is used in the context analysis step described in Section 2.2.

The resulting ranked context lists demonstrate the different contexts in which words like "Iraq" and "rock" appear in the data. The top 5 context words for a window size of 20 for "Iraq" are *inge, chirac's, refusal, reconstruction,* and *waging.* The top 5 context words for "rock" are *uplifting, kt, folk, lejeune,* and *assertion.* Most of the top words in the first list are, for the most part, relevant to the word "Iraq," and the words in the second list are clearly not relevant to "Iraq." The corresponding top 5 list for "Abbas" is *mahmoud, ariel, prime, minister,* and *committed;* the list for "bus" is *michelle, blew, moscow, jerusalem,* and *responsible.*

These lists also demonstrate the value of modeling the patterns in the ASR output directly, rather than compiling co-occurrence frequencies from a clean data source without word errors: the output word *inge* occurs exclusively in the data as an ASR error for *in going* in the context "in going to war with Iraq." Similarly, *kt* occurs frequently in the data as part of the call letters for a television station in Little Rock, Arkansas. Systematic and recurring errors such as this provide a great deal of information in the co-occurrence statistics. However, the use of ASR output without "clean" transcripts in training also introduces the possibility of modeling false positives in the

output, such as "Iraq" being output as an error when "a rock" was spoken; this type of error can adversely affect the co-occurrence statistics we calculate.

2.2 Context Analysis

The context analysis component seeks to identify contextual regions in the test data that are likely to contain a given query word, and thus also likely to contain a misrecognition of the query word. This analysis uses the probabilities and mutual information output from the co-occurrence analysis described in Section 2.1.

We slide a window of w words across a document in the test data, where w is the same window size used to train the word co-occurrence statistics. We also define a minimum number of context words c that must be contained with the window in order to mark the center word of the window as a possible ASR error. As an example of this context matching, consider the word sequence "... the reconstruction of a rock proceeds despite Chirac's refusal..." The word "rock" is at the center of an 8-word context window (4 on either side) containing 3 of the top-ranked context words for "Iraq" from the previous section. This instance of the word "rock" would thus be a candidate misrecognition of "Iraq" for $w \geq 8$ and $c \leq 3$.

Table 1 shows the number of candidate words detected for "Iraq" in the evaluation data for different window sizes w and minimum context words c. As might be expected, the number of candidate words increases as the window size increases and decreases as the minimum number of context words increases.

	Window Size w				
c	2	6	10	14	20
1	10412	27643	42142	54507	70212
2	346	3941	8820	14409	23133
3	x	597	2152	4314	8418
4	x	84	505	1411	3437
5	x	9	123	438	1387

Table 1: Candidate errors for "Iraq" detected within a range of window sizes w for minimum numbers of context words c.

Most combinations result in a large number of candidates, so we also apply candidate pruning based on probabilistic metrics. Given a candidate error and c context words contained in a window, we then compare the probability of observing both the query word and the actual word in the data. This comparison is carried out using the Kullback-Leibler divergence for observation distributions containing the c context words, $D(p \parallel q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)}$, where $p(x)$ is the conditional probability of the co-occurrence of the query word with a context word x in the set of c context words X and $q(x)$ is the probability of the candidate error with the context word. A larger Kullback-Leibler divergence value indicates a higher probability that the candidate word is actually a misrecognition of the query word.

2.3 Phonetic Comparison

Given the set of candidate errors for a query word, as detected using the context matching technique described in Section 2.2, we next apply a phonetic distance criterion to determine the similarity between each candidate error and the query word being sought, based on the pronunciations in the ASR system lexicon. We used the common minimum-distance weighted phonetic alignment technique described in detail in (Kondrak, 2003); in our experiments we used phonetic weights available through the alignment package *altdistsm* originally described in (Fisher and Fiscus, 1993).

The final decision whether to correct the ASR error is made based on the phonetic distance between the candidate word and the query word. Since the candidate word is already known to have occurred in a lexical context that is likely to contain the query, a strong phonetic similarity between the words provides very strong evidence that the candidate word is, in fact, a misrecognition of the query word.

3 Results and Discussion

We carried out an initial evaluation of our system using three specific query words that were featured in a large number of news stories in the training corpus: "Iraq," "Abbas," and "Lynch" (from Jessica Lynch, an American soldier during the war in Iraq). The 39 files in the test corpus were annotated to indicate all the locations of recognition errors involving these three spoken words. In addition, the location of errors that are morphological variants of the query word, such as "Iraqi" and "Iraq's" were annotated and were not included in the evaluation results; in the context of information retrieval these morphological variants can easily be addressed using common techniques such as stemming.

The query word "Lynch" turned out to be an uninteresting case for our approach: it was misrecognized only 4 times in the test corpus, each time as the morphological variant "lynched." Nevertheless, the context matching test worked well, as three of the top-ranked context words were the very relevant "Jessica," "private," and "rescue." The detection and correction results for the word "Abbas" were also very encouraging, although the small sample size makes it difficult to draw significant conclusions. In our test corpus, there were n=10 examples in which "Abbas" was misrecognized. Our method detected 8 candidates, 7 of which were actually misrecognitions of "Abbas," for a recall of 70% and a precision of 88% (window size w=10, minimum context c=2). Corrections included

"a bus," "a bass," and "a boss," and the false positive was the word "about," which is phonetically very similar.

The query term "Iraq" proved to be the most fruitful query term, due to its prevalence throughout the corpus. There was a total of 142 cases in which "Iraq" was misrecognized. Examples of common errors were "rock," "a rock," "your rocks," "warren rock" (war in Iraq), "her rock," "any rocket" (in Iraq), and "a rack." Table 2 shows the final results for the query term "Iraq" for the 39 ASR output test files, for a range of minimum required context words c and the most-successful window size (14).

c	Detect	Correct	False Pos	R	P
1	138	120	18	85	87
2	92	87	5	61	95
3	51	51	0	36	100
4	27	27	0	19	100
5	9	9	0	6	100

Table 2: Results for query word "Iraq" for window size w=14 and a range of minimum context words c: hypothesized errors detected and corrected, false positives, recall, and precision (n=142).

Although we can not draw conclusions about the general applicability of this approach until we carry out further experiments with more test cases, the preliminary detection and correction results indicate that it is possible to achieve very high precision with a reasonable recall for certain window sizes and numbers of context words. Table 3 shows recall and precision values for some of the most effective combinations of window sizes w and minimum context words c which return few false positives and many accurate corrections.

w	c	R	P
8	1	69	88
8	2	38	100
10	1	77	87
10	2	44	98
14	2	61	95
14	3	36	100

Table 3: Recall and precision values for selected window and minimum context values.

The work we describe in this paper is complementary to ASR algorithmic improvements, in that we treat error detection and correction as a post-processing step that can be applied to the output of any ASR system and can be adapted to incremental improvements in the systems. This form of post-processing also allows us to take advantage of long-range contextual features that are not available during the ASR decoding itself. Post-processing also enables large-scale data analysis that models the types of systematic errors that ASR systems make. All the steps in our approach, co-occurrence analysis, context matching, and phonetic distance pruning, are unsupervised methods that can be automatically run for large quantities of data.

The results in this paper are promising but are obviously very preliminary. We are in the process of evaluating the work on a much larger set of query words. We should emphasize that the goal of this work is not to produce a significant improvement in the overall word error rate of a particular corpus of ASR output, although we believe that such an improvement is possible using similar contextual analysis. Instead, the focus of the work is to improve the specific aspects of the ASR output that may adversely affect a user-centered task like information retrieval. While we have not formally evaluated the impact of our error detection and correction on retrieval performance, there is an obvious benefit to correcting misrecognitions of the specific query term that a user is seeking.

References

Jerome R. Bellegarda, "A multi-span language modeling framework for large vocabulary speech recognition," *IEEE Transactions on Speech and Audio Processing*, 6:456–467, 1998.

W. M. Fisher and J.G. Fiscus, "Better Alignment Procedures for Speech Recognition Evaluation," *Proc. International Conference on Acoustic, Speech and Signal Processing*, pp. II-59 - II-62, 1993.

G. Kondrak, "Phonetic Alignment and Similarity," *Computers and the Humanities* 37(3), August 2003, pp. 273–291.

J. Kupiec, D. Kimber, and V. Balasubramanian, "Speech-Based Retrieval Using Semantic Co-Occurrence Filtering," In *Proc. Human Language Technologies*, pp. 373-377, 1994.

B. Logan and J.M. Van Thong, "Confusion-based Query Expansion for OOV Words in Spoken Document Retrieval," In *Proc. ICSLP 2002,*, Denver, Colorado, pp. 1997-2000, 2002.

J. Makhoul, F. Kubala, T. Leek, D. Liu, L. Nguyen, R. Schwartz, and A. Srivastava, "Speech and Language Technologies for Audio Indexing and retrieval," In *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1338-1353, 2000.

B. Roark and E. Charniak, "Noun-Phrase Co-Occurrence Statistics for Semi-Automatic Semantic Lexicon Construction," In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics*, pp. 1110-1116, 1998.

R. Rosenfeld, "Two decades of Statistical Language Modeling: Where Do We Go From Here?" *Proceedings of the IEEE*, 88(8), 2000.