Error Detection in Automatic Speech Recognition

Farshid Zavareh, Ingrid Zukerman, Su Nam Kim and Thomas Kleinbauer
Faculty of Information Technology, Monash University
Clayton, VICTORIA 3800, Australia
sfhos2@student.monash.edu,
{Ingrid.Zukerman,Su.Kim,Thomas.Kleinbauer}@monash.edu

Abstract

We offer a supervised machine learning approach for recognizing erroneous words in the output of a speech recognizer. We have investigated several sets of features combined with two word configurations, and compared the performance of two classifiers: Decision Trees and Naïve Bayes. Evaluation was performed on a corpus of 400 spoken referring expressions, with Decision Trees yielding a high recognition accuracy.

1 Introduction

One of the main stumbling blocks for spoken Natural Language Understanding (NLU) systems is the lack of reliability of Automatic Speech Recognizers (ASRs) (Pellegrini and Trancoso, 2010). Recent research prototypes of ASRs yield Word Error Rates (WERs) between 15.6% (Pellegrini and Trancoso, 2010) and 18.7% (Sainath et al., 2011) for broadcast news. However, the WER of the ASR we employed (Microsoft Speech SDK 6.1) is 34% when trained on an open vocabulary plus a small language model for our corpus. This WER is consistent with that obtained in the 2010 Spoken Dialogue Challenge (Black et al., 2011).

In this paper, we offer a supervised machine learning approach to detect erroneous words in ASR output (this step will be followed by automatic error correction). Our approach was evaluated on a corpus of 400 spoken referring expressions, with the best-performing option yielding an average accuracy of 89% (Section 5).

The rest of this paper is organized as follows. In the next section, we discuss related work. In Section 4, we describe our experimental design, focusing on the features considered for our machine-learning approach. In Section 5, we discuss our results, followed by concluding remarks.

2 Related Research

Approaches for improving the performance of spoken NLU systems may be classified into *prevention* and *recovery*.

Prevention avoids errors by constraining the vocabulary (Gorniak and Roy, 2005; Sugiura et al., 2009) and grammatical constructs (Brooks and Breazeal, 2006) understood by an ASR. ASRs that employ this approach can process expected utterances efficiently, and work well in restricted domains. However, these ASRs have trouble processing unexpected utterances.

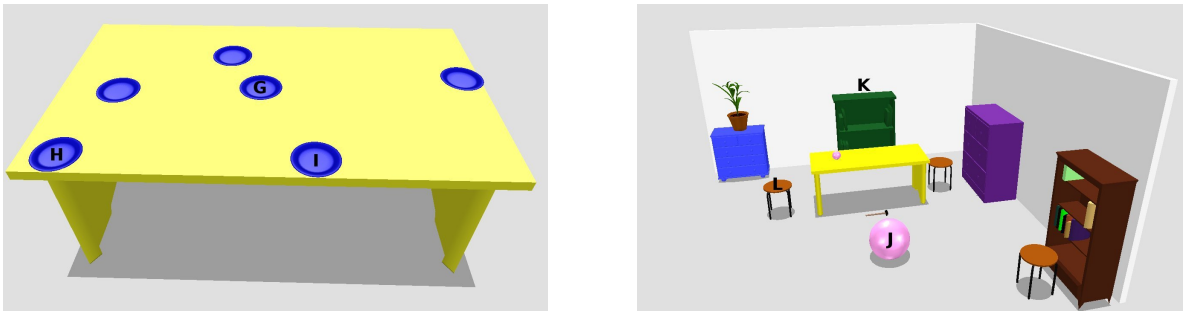
Recovery involves *error detection* followed by *correction*. During detection, an NLU system posits that a word in an utterance was incorrectly recognized. Three approaches to error recovery are described in (López-Cózar and Griol, 2010; Ringger and Allen, 1996; Zhou et al., 2006).

López-Cózar and Griol (2010) consider statistical information, and lexical, syntactic, semantic and dialogue-related information to correct ASR errors (i.e., replace, insert or delete words in a textual ASR output), and syntactic approaches to modify tenses of verbs and grammatical numbers to better match grammatical expectations.

Ringger and Allen (1996) use statistical information to construct a language model that quantifies the likelihood of word sequences, and a noisy channel model that predicts errors made by an ASR. They perform error detection and correction at the same time based on these models, which are trained using the words expected in the domain.

Zhou *et al.* (2006) perform error detection and correction of utterances, words and characters in Mandarin. They experiment with the *Generalized Word Posterior Probability* (*GWPP*) of an utterance, computed from word hypotheses, utterance length, language model, and acoustic observations; and features based on the N-best hypotheses, obtained from acoustic, language model and purity scores. When an erroneous word is de-

Farshid Zavareh, Ingrid Zukerman, Su Nam Kim and Thomas Kleinbauer. 2013. Error Detection in Automatic Speech Recognition. In *Proceedings of Australasian Language Technology Association Workshop*, pages 101−105.



(a) Projective relations and "end, edge, corner" and "center" of a table

(b) Colour, size, positional relation and intervening object in a room

Figure 1: Two of the scenarios used to construct our corpus.

tected, all the characters in it are deemed to be wrong. Correction is then performed using a list of candidate alternatives for each erroneous character to generate a list of word hypotheses, and a linguistic model based on mutual information and trigrams to select the best word hypothesis.

Like these researchers, we offer corpus-based techniques to detect ASR errors. However, we employ features of the ASR output, rather than actual words or expectations from the context. By doing this, we hope to avoid over-fitting to domain-specific words and expectations.

3 The Corpus

Error detection performance was evaluated using the corpus constructed by Kleinbauer *et al.* (2013). The corpus originally comprised 432 free-form descriptions spoken by 26 trial subjects to refer to 12 designated objects in four scenarios (three objects per scenario, where a scenario contains between 8 and 16 objects; two scenarios appear in Figure 1). Half of the participants were native English speakers, and half were non-native. All the speakers were proficient in English, but the non-native speakers had a foreign accent, and some had idiosyncratic turns of phrase.

We manually filtered out 32 descriptions that were broken up by the ASR due to pauses made by the speakers, leaving 400 descriptions, which comprise 3,128 words in total, and 118 unique words. The descriptions, which varied in length and complexity, had an average length of 10 words and a median length of 8 words, with the longest description containing 21 words. Sample descriptions are: "the green plate next to the screwdriver at the top of the table", "the large pink ball in the middle of the room", "the plate in the corner of the table", and "the picture on the wall".

The ASR produced up to 50 alternative textual interpretations for each spoken description, ranked in descending order of probability. In total, 4,249 texts, with 33,927 words (706 unique) were generated. It is worth noting that more alternatives, with a higher average WER for the top-ranked options, were generated for non-native speakers than for native speakers.

We used the Levenshtein distance to align each alternative produced by the ASR with the reference (correct) description. The words in the alternative were then labeled as follows: **C**orrect, **I**nserted – absent from the reference interpretation, **R**eplaced – an incorrect word instead of the reference word, and **D**eleted – a placeholder for a reference word that is not in the alternative. The Inserted and Replaced words comprise the *Wrong* class (Deleted words cannot be modeled).

4 Experimental Design

In this section, we discuss the classifiers we considered, our feature sets, and evaluation methods.

Classifiers. We investigated two classifiers to decide whether a word in a text produced by the ASR is correct: Decision Trees (DT) (Quinlan, 1993) and Naïve Bayes classifiers (NB) (Domingos and Pazzani, 1997) (`cs.waikato.ac.nz/ml/weka/`).[1] For NB, we used equal-width binning to discretize continuous features (Catlett, 1991; Kerber, 1992).

Features. The target classes are *Correct* or *Wrong*, and three types of features were computed for each word w in a text: word based (5), sentence based (6), and phoneme based (2).

Word-based features. (1) *Part of Speech (PoS)* as determined by the Stanford PoS Tagger

[1] Initially we also considered linear chain Conditional Random Fields (CRF) (Lafferty et al., 2001) (`mallet.cs.umass.edu`), but they exhibited inferior performance.

(nlp.stanford.edu/software/tagger.shtml); (2) *Stop Word* as determined by the list in webconfs.com/stop-words.php; (3) *Position* of w in the text, defined as a nominal feature taking one of the values **B**eginning, **M**iddle or **E**nd; (4) *Time* taken by the speaker to pronounce word w (in fraction of a second); and (5) *Confidence Score* given to word w by the ASR.

Sentence-based features. (6) *Repetition Count* – number of alternatives where w is repeated; (7) *Repetition Ratio* (equivalent to purity score (Zhou et al., 2006)) – *Repetition Count* divided by the total number of alternatives; (8) *Replacement Ratio* – number of alternatives which, when aligned with the current alternative, label w with "R", divided by the total number of alternatives; (9) *Insertion Ratio* – number of alternatives which, when aligned with the current one, label w with "I", divided by the total number of alternatives; (10) *Rank* of the alternative containing w in the ASR output; and (11) *Sentence Length* – number of words in the current alternative.

Phoneme-based features (according to the CMU Pronunciation Dictionary, speech.cs.cmu.edu/cgi-bin/cmudict)**.** (12) *Broad Sound Groups* (*BSGs*) – a vector of length 8 that represents the number of times each BSG occurs in word w, e.g., the word "problem" has 2 vowels, 2 stops, 2 liquids, and 1 nasal; and (13) *Phonemes* – a vector of length 39 that represents the number of times a phonetic symbol appears in w's phonetic transcription.

We experimented with the following sets of features: (1) *Word + Sentence* features, (2) *BSGs*, and (3) *Phonemes*. These features were computed for the current word (C), which is being classified, and for the previous, current and next word (*PCN*). For example, the following vector is produced when all 58 features are used for the current word (the first and last word in an alternative have missing features for P and N respectively):

$$\underbrace{f_1,\ldots,f_5}_{\textit{Word}},\underbrace{f_6,\ldots,f_{11}}_{\textit{Sentence}},\underbrace{f_{12},\ldots,f_{19}}_{\textit{BSGs}},\underbrace{f_{20},\ldots,f_{58}}_{\textit{Phonemes}}.$$

Sets of features that included actual words produced accuracies of over 95%, but were unlikely to generalize. This was evident by inspecting the generated decision tree, which was shallow and wide. In fact, when w was used, most other features were ignored. Consequently, we decided not to include the actual words in our feature sets.

Table 1: Accuracy of DT versus NB: Different feature combinations.

Classifier	Features	Micro-average	Macro-average
NB	*Word+Sentence*, C	0.8156	0.8146
NB	*Word+Sentence*, *PCN*	0.8060	0.8066
NB	*BSGs*, C	0.6479	0.6446
NB	*BSGs*, *PCN*	0.6476	0.6479
NB	*Phonemes*, C	0.6610	0.6605
NB	*Phonemes*, *PCN*	0.6722	0.6731
DT	*Word+Sentence*, C	0.8110	0.8110
DT	*Word+Sentence*, *PCN*	0.8082	0.8121
DT	*BSGs*, C	0.7959	0.7974
DT	*BSGs*, *PCN*	0.8308	0.8324
DT	*Phonemes*, C	0.8614	0.8591
DT	***Phonemes, PCN***	**0.8771**	**0.8770**

Evaluation method. We employed 13-fold cross validation to train and test our corpus, where each fold comprises descriptions spoken by one native English speaker and one non-native speaker (Section 3). The per-speaker split ensures that sentences spoken by one trial subject do not appear in both training and test sets; and the native/non-native pairing balances the test sets, in the sense that they are of similar size, and ASR performance is similar for all sets (Section 3).

5 Results

Table 1 shows the results of our initial tests, which compare the performance of DT with that of NB in terms of micro- and macro-averaged accuracy (recall that the majority class of *Correct* words is 66%, Section 1). The odd-numbered rows contain the results for the three sets of features computed only for C, and the even-numbered rows contain the results for *PCN*. The statistically significant best result is boldfaced (statistical significance was calculated using the Paired Student's t-test).

As seen in Table 1, compared to C, *PCN* has a mixed effect on NB's performance, depending on the base features: *PCN* yields a statistically significant drop in accuracy for *Word+Sentence* (*p-value*=0.03), no statistically significant change for *BSGs*, and an improvement for *Phonemes* (*p-value*=0.015). The results are more consistent for DT: there is no significant difference in performance between C and *PCN* for *Word + Sentence*, but *PCN* yields statistically significant improvements for the other feature sets (*p-value* ≤ 0.05).

There were no statistically significant differences in accuracy between DT and NB for *Word+Sentence* with C and *PCN*. However, DT significantly outperformed NB in the remaining tests (*p-values* $<<$ 0.01). In addition, *PCN*

Table 2: Accuracy comparison for DT with *Phonemes* plus different feature combinations.

Features ***Phonemes, PCN +***	**Micro-average**	**Macro-average**
	0.8771	0.8770
Word+Sentence	0.8775	0.8787
BSGs	0.8776	0.8783
Word+Sentence and BSGs	0.8741	0.8754
PoS	**0.8902**	**0.8906**
PoS and BSGs	**0.8972**	**0.8971**

yielded a better performance than C for DT. Hence, our next tests are carried out using DT with *PCN* only.

Table 2 shows the results of combining *Phonemes*, which give the best accuracy (Table 1), with three feature sets: *Word+Sentence*, *BSGs* and *PoS*. The last two rows in Table 2 (boldfaced) show the feature sets that yield the highest (statistically equivalent) accuracies. These results, which were obtained with *PoS*, with and without *BSGs*, are significantly better than those achieved when *Word + Sentence* features or *BSGs* were used (*p-value* $\leq$ 0.05). Also, combining *Phonemes* with *Word+Sentence*, *BSGs* and both *Word+Sentence* and *BSGs* does not yield significant performance changes.

The most significant features in the best-performing decision trees are (in descending order): presence of the phonemes TH and Z, number of occurrences of N ($\leq$1 versus 1$<$), whether *PoS*=JJ (adjective), and whether the next word contains a stop *BSG* (at level 5 in the tree). This indicates that certain phonemes are prone to ASR mis-interpretation — an insight that has significant implications for the next stage of the ASR process, which consists of proposing replacements for words that are classified as *Wrong*. For example, we could create a confusion matrix between error-prone phonemes produced by the ASR and likely replacement phonemes, and suggest replacement words that include these hypothesized phonemes (Thomas et al., 1997; Zhou et al., 2006). It is worth noting that the ASR's *Confidence Score* was not used in the best-performing DTs. In fact, we observed that this score is often inconsistent with the *Correct/Wrong* class of a word.

As mentioned in Section 4, using the actual words as a classification feature yielded decision trees that over-fitted the data. Thus, it is possible that a similar effect takes place when *Phonemes* are used. Additional tests on different datasets should be conducted to rule out this possibility. Notice, however, that *BSGs* with *PCN* yield a creditable performance (third last row in Table 1), which improves statistically significantly (*p-value* $<<$ 0.01) when *BSGs* are combined with *PoS* and *Word+Sentence* (Table 3). This is noteworthy because *BSGs* are abstractions of *Phonemes*, and hence are less likely than *Phonemes* to fit a small number of words. Further, a correction procedure similar to that suggested for *Phonemes* would be applicable for *BSGs*.

Table 3: Accuracy comparison for DT with *BSGs* plus different feature combinations.

Features ***BSGs, PCN +***	**Micro-average**	**Macro-average**
	0.8308	0.8324
Word+Sentence	**0.8640**	**0.8626**
PoS	**0.8639**	**0.8632**

6 Conclusions and Future Work

We have proposed a supervised learning method to predict the correctness of words in an ASR output. Our best classifier yields 89% accuracy. However, these results were obtained on a relatively small corpus with a limited vocabulary (Section 3). Hence, further tests with larger, more diverse corpora are needed to verify our results.

As mentioned in Section 3, we aligned the alternatives returned by the ASR with the reference text in order to label the words in each alternative. In addition, we aligned the alternatives with each other to compute multi-alternative features, such as *Repetition count* and *Replacement ratio*. In doing so, we implicitly assumed that there is a one-to-one mapping between the words in an alternative and those in the reference text, and also between the words in alternatives generated for the same spoken description. However this assumption is not always valid: we have observed cases where one word has been split into two words by the ASR, or a few words have been merged into one. Ringger and Allen (1996) have proposed a statistical solution to this problem, but unfortunately their method relies heavily on the vocabulary on which the system was trained. This problem will be addressed in the future.

The methods offered in this paper do not distinguish between a *Wrong* word and *Noise* (sighs or hesitations that are often mis-heard by the ASR as "and", "on" or "in"). In the future, we propose to retrain our system to deal with three classes, viz *Correct*, *Wrong* and *Noise*.

Acknowledgments

This research was supported in part by grants DP110100500 and DP120100103 from the Australian Research Council. The authors thank Masud Moshtaghi for his help with statistical issues.

References

A. Black, S. Burger, A. Conkie, H. Hastie, S. Keizer, O. Lemon, N. Merigaud, G. Parent, G. Schubiner, B. Thomson, J.D. Williams, K. Yu, S. Young, and M. Eskenazi. 2011. Spoken dialog challenge 2010: Comparison of live and control test results. In *Proceedings of the 11th SIGdial Conference on Discourse and Dialogue*, pages 2–7, Portland, Oregon.

A.G. Brooks and C. Breazeal. 2006. Working with robots and objects: Revisiting deictic reference for achieving spatial common ground. In *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-robot Interaction*, pages 297–304, Salt Lake City, Utah.

J. Catlett. 1991. On changing continuous attributes into ordered discrete attributes. In *EWSL-91 – Proceedings of the European Working Session on Learning*, pages 164–178, Porto, Portugal.

P. Domingos and M. Pazzani. 1997. On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, 29:103–130.

P. Gorniak and D. Roy. 2005. Probabilistic grounding of situated speech using plan recognition and reference resolution. In *ICMI'05 – Proceedings of the 7th International Conference on Multimodal Interfaces*, pages 138–143, Trento, Italy.

R. Kerber. 1992. ChiMerge: Discretization of numeric attributes. In *AAAI92 – Proceedings of the 10th National Conference on Artificial Intelligence*, pages 123–128, San Jose, California.

Th. Kleinbauer, I. Zukerman, and S.N. Kim. 2013. Evaluation of the *Scusi?* spoken language interpretation system – A case study. In *IJCNLP2013 – Proceedings of the 6th International Joint Conference on Natural Language Processing*, pages 225–233, Nagoya, Japan.

J.D. Lafferty, A. McCallum, and F.C.N. Pereira. 2001. Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data. In *ICML'2001 – Proceedings of the 18th International Conference on Machine Learning*, pages 282–289, Williamstown, Massachusetts.

R. López-Cózar and D. Griol. 2010. New technique to enhance the performance of spoken dialogue systems based on dialogue states-dependent language models and grammatical rules. In *Proceedings of Interspeech 2010*, pages 2998–3001, Makuhari, Japan.

T. Pellegrini and I. Trancoso. 2010. Improving ASR error detection with non-decoder based features. In *Proceedings of Interspeech 2010*, pages 1950–1953, Makuhari, Japan.

J. R. Quinlan. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, California.

E. Ringger and J.F. Allen. 1996. A fertility channel model for post-correction of continuous speech recognition. In *ICSLP-96 – Proceedings of the 4th International Conference on Spoken Language Processing*, pages 897–900, Philadelphia, Pennsylvania.

T.N. Sainath, B. Ramabhadran, M. Picheny, D. Nahamoo, and D. Kanevsky. 2011. Exemplar-based sparse representation features: From TIMIT to LVCSR. *IEEE Transactions on Audio, Speech and Language Processing*, 19(8):2598–2613.

K. Sugiura, N. Iwahashi, H. Kashioka, and S. Nakamura. 2009. Bayesian learning of confidence measure function for generation of utterances and motions in object manipulation dialogue task. In *Proceedings of Interspeech 2009*, pages 2483–2486, Brighton, United Kingdom.

I.E. Thomas, I. Zukerman, I. Oliver, D. Albrecht, and B. Raskutti. 1997. Lexical access for speech understanding using Minimum Message Length encoding. In *UAI'97 – Proceedings of the 13th Annual Conference on Uncertainty in Artificial Intelligence*, pages 464–471, Providence, Rhode Island.

Z. Zhou, H.M. Meng, and W.K. Lo. 2006. A multi-pass error detection and correction framework for Mandarin LVCSR. In *Proceedings of Interspeech 2006*, pages 17–21, Pittsburgh, Pennsylvania.