ERROR CORRECTIVE MECHANISMS FOR SPEECH RECOGNITION

Lidia Mangu and Mukund Padmanabhan

IBM T.J. Watson Research Center, P.O. Box 218,
Yorktown Heights, NY 10598
{mangu,mukund}@us.ibm.com

ABSTRACT

In the standard MAP approach to speech recognition, the goal is to find the word sequence with the highest posterior probability given the acoustic observation. Recently, a number of alternate approaches have been proposed for directly optimizing the word error rate, the most commonly used evaluation criterion. One of them, the consensus decoding approach, converts a word lattice into a confusion network which specifies the word-level confusions at different time intervals, and outputs the word with the highest posterior probability from each word confusion set. This paper presents a method for discriminating between the correct and alternate hypotheses in a confusion set using additional knowledge sources extracted from the confusion networks. We use transformation-based learning for inducing a set of rules to guide a better decision between the top two candidates with the highest posterior probabilities in each confusion set. The choice of this learning method is motivated by the perspicuous representation of the rules induced, which can provide insight into the cause of the errors of a speech recognizer. In experiments on the Switchboard corpus, we show significant improvements over the consensus decoding approach.

1. INTRODUCTION

Most state-of-the-art speech recognizers output word lattices as a compact representation of a set of alternate sentence hypotheses. The output of the standard Viterbi decoding algorithm is the path in the word lattice with the highest likelihood. Alternative decoding methods for directly minimizing word error rate have been proposed recently [4,8,11]. In the consensus decoding approach [8], the lattice is converted into a confusion network which specifies a sequence of word confusions and the posterior probability of each word. The decoded output is then formed by concatenating the words with the highest posterior probabilities in this sequence. The motivation for this paper comes from the observation that a significant number of times, the candidate with the second highest posterior probability in a confusion set is the correct one, rather than the candidate with the highest posterior probability. This paper is a first attempt towards finding a better algorithm for selecting between the best two candidates of a confusion set. Our goal is to *automatically* infer a set of rules for deciding when to prefer the second candidate over the first one. The features used in the learner are context-independent, based entirely on the properties of the confusion set to be disambiguated. It is part of future work to investigate the usefulness of neighbouring context in the decision-making process.

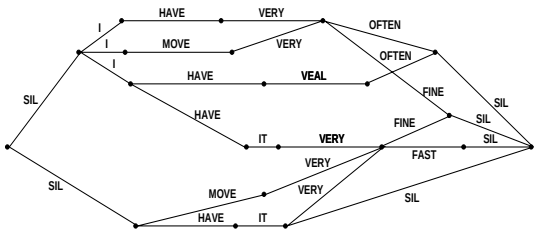
Posed as a confusion set disambiguation task, this problem becomes very similar to the spelling correction task as described in [5,9]. In [9], it is shown that transformation-based learning used for spelling correction not only achieves good results in terms of accuracy, but also provides a perspicuous representation for the acquired knowledge.

The paper is organized as follows. In Sections 2 and 3, we describe the basic principles of the consensus decoding and transformation-based learning paradigms. In Section 4, we present a transformation-based system that learns rules for correcting the output of consensus decoding, and describe some experiments and their results. We present conclusions and discuss future work in Section 5.

2. CONSENSUS DECODING

In the standard approach to speech recognition, the goal is to find the sentence hypothesis W with the highest posterior probability $P(W|A)$ given the acoustic observation A. This is equivalent to minimizing the *sentence* error rate. The most commonly used evaluation metric is *word* error rate (WER); hence, there is a mismatch between the decoding and evaluation criteria. The solution proposed in the consensus decoding framework [8] has the advantage that, in addition to minimizing WER, it also produces a new representation of the set of candidate hypotheses that specifies the sequence of word level confusions. Starting with a word lattice, it first groups together the links corresponding to the same word instance. The similarity metric used is based on time information associated with the links. Then, it merges heterogeneous sets of links based on the phonetic similarity

(a) Input lattice



(b) Confusion Network ("-" marks deletions)

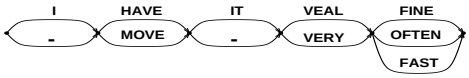


Fig. 1. From lattices to confusion networks

of the word components. The result of the clustering process is a sequence of *confusion sets*. Each set contains all of the words that are alternates in a particular time interval. The posterior probability of a word in a confusion set is computed by summing over the posterior probabilities of all of the links associated with the word. The posterior probability of a deletion ("-") in a confusion set is the remaining posterior probability mass not assigned to the words in the set. The *consensus hypothesis* is obtained by concatenating the words with the highest posterior probabilities in the sequence of confusion sets. The *confusion network* is the graphical representation obtained at the end of the clustering procedure. Figure 1 illustrates an example of a lattice and the corresponding confusion network.

It has been shown that the consensus hypothesis results in consistent improvements over the MAP hypothesis on a variety of tasks [3,6,7,8]. This paper presents a method for improving the consensus hypothesis by making use of the additional information existing in confusion networks. Reducing a lattice to a confusion network allows us to replace the global search over a large set of sentence hypotheses with a local search over a small set of word hypotheses. This feature allows us to cast decoding as a classification problem and approach it with standard machine learning techniques.

3. TRANSFORMATION-BASED LEARNING

Transformation-based learning has been applied successfully to a number of natural language problems, including part-of-speech tagging, prepositional phrase attachment, parsing and spelling correction, often achieving state-of-the-art accuracy while capturing the acquired knowledge in a small set of rules [1,9].

To fully specify a transformational system, one must specify a baseline predictor, a set of allowable transformation types and an objective function for learning. In learning, the training set is first annotated based on some baseline predictor and the goal of the learner is to learn a sequence of corrective rules. A single iteration of the learner consists of the following steps. First, apply each possible transformation to a copy of the current training corpus and score it based on the objective function. Second, pick the rule with the highest score, append it to the end of the transformation list and apply it to the training set.

The result of learning is an ordered list of transformations. In testing, first apply the initial predictor to the test set and then apply each rule, in order, everywhere it can be applied. It has been shown that, for a fixed set of features, transformation lists are more powerful than decision trees in what they can learn [1].

4. TRANSFORMATION-BASED LEARNING IN THE CONSENSUS DECODING FRAMEWORK

In this section we describe how to build a transformation-based system for discriminating between the two highest posterior probability words in a confusion set. The baseline predictor initially assumes the highest ranked candidate is the correct one. The allowable transformations in these experiments are described by the following template:

- Change c_1 to c_2 if
  $A_1 op_1 v_1$ and $A_2 op_2 v_2$ and .... $A_k op_k v_k$.

where $op_i \in \{=, <, >\}$, A_i are the features extracted from each confusion set, having categorical or integer values v_i, and $c_1, c_2 \in 1, 2$ correspond to choosing the first or the second candidate, respectively. The features used by the learner in this paper are:

- Word identity, duration and posterior probability of the two competing words
- Difference in the posterior probabilities of the two candidates
- Temporal position of the confusion set in the sentence
- Number of candidates in the confusion set

For example, one rule that could be learned is "Choose the second candidate if the first candidate is the word "A", the second candidate is "-"(deletion) and the difference in posterior probabilities between the two is less than 0.1". This type of rule would not be surprising, since most recognizers tend to insert short words.

The objective function used in this experiment is the classification accuracy, which is directly correlated to word error rate.[1] At each confusion set where the current choice is incorrect, the rule templates are used to form candidate

[1]The difference between the classification accuracy and WER comes from different methods of aligning a hypothesis to the reference; in one case the alignment is done via confusion networks.

rules for correction. We identify all the rules that would have a *positive* effect on the current confusion set. By testing them against the rest of the training set, we obtain a count of the negative effects each rule has. Each rule is assigned a score based on the number of positive and negative changes caused by applying the rule.

In the standard transformation-based learning approach, the iterative process continues until no transformation can be found whose application results in an improvement to the training corpus. In our experiments, the best stopping rule found was based on statistical significance. Therefore, rules with low scores were not considered unless they were statistically significant[2].

4.1. Experimental Setup

The speech recognition system used in our experiments is an HMM-based state-clustered, context-dependent system, containing 3140 states and 277K diagonal covariance prototypes. The system uses 60-dim HDA+MLLT features [10]. The language model is a trigram backoff model trained on Switchboard, Broadcast News and Callhome data.

A prerequisite for the success of our approach is that the same error patterns are observed in training and testing. Hence, we can not use a system which was trained using the entire acoustic training data, to decode utterances extracted from the same data. Therefore, we built two systems, *Small* and *Big*, trained on 60 hours and 243 hours of acoustic data, respectively. The purpose of building a *Big* system is to check if the rules induced for correcting the *Small* system can still be applied successfully on the *Big* one.

From the acoustic data not used in training system *Small*, we extracted 4000 utterances for rule training and 2000 utterances as the held-out set. The held-out set was used just for validating the statistics extracted from the training set. The test set consists of 2427 utterances from 19 conversations comprising 18000 words. This set was used in the 1997 Johns Hopkins University LVCSR Workshop (WS97).

We used system *Small* to produce the confusion networks for the rule training and the held-out set and ranked the words in each confusion set based on their posterior probability. We then aligned the correct transcriptions to the confusion networks (with a simple dynamic programming procedure) and computed the percentage of time the correct word was at a specified rank. The results are shown in Table 1. There are two things worth observing in this table. First, the statistics are very similar for both the training and the held-out set. Second, most of the potential accuracy improvement over the consensus hypothesis comes from the second-ranked candidate. Table 2 shows the potential WER improvement if we are able to select perfectly between the top k candidates. It can be seen that there is around 13% absolute WER reduction if an oracle picks between the top 2 candidates in a confusion set.

[2]We used the likelihood ratio test described in [2].

Rank	Classification accuracy (%)	
	Training set	Held-out set
1	73.3	73.2
2	10.3	10.8
3	3.8	3.9

Table 1. Accuracy for different ranks in the confusion sets.

k	Oracle WER (%)	
	Training set	Held-out set
1	38.0	37.5
2	25.1	24.5
3	20.0	19.3

Table 2. Oracle WER when limiting the choice to the k words with the highest posterior probability in the confusion sets.

For our training procedure we need confusion sets containing at least two candidates. Among all the confusion sets in the training set, 23% contain only one candidate and this word is correct 95% of the time.

We built an initial training set from all of the confusion sets in which the correct word is either the first or the second word. By examining this set, we found that when the highest ranked word has a posterior probability greater than 0.8, this word is correct in more than 92% of the cases. We consider this baseline accuracy hard to improve upon; therefore, we restrict ourselves only to the cases in which the correct word is one of the two top candidates, and the posterior of the highest ranked one is less than 0.8. This final training set contains 23% of all confusion sets and the top word has a baseline classification accuracy of 67%. The potential overall WER improvement if we always choose correctly between the first and second candidate in the confusion sets in this subset, is around 10% absolute.

4.2. Results

We applied the greedy learning stategy described in Section 4, which proceeded by searching over a span of a few thousand rules and produced a list containing 10 transformational rules. For testing those rules, we first created the confusion networks on the test set using the two speech recognition systems *Small* and *Big* described above. We used the standard consensus decoding for confusion sets with the highest posterior probability greater than 0.8 and for the ones containing only one candidate, and applied the learned rules only on the remaining confusion sets. The new hypothesis, which we will refer to as *consensus+*, is formed by concatenating the words predicted in each confusion set in the order specified by the confusion network. Table 3

Hypothesis	Word Error Rate (%) WS97 (*Small*)	WS97 (*Big*)
MAP	38.0	36.0
Consensus	37.2	35.1
Consensus+	36.4	34.6

Table 3. Comparison of the WER results for the consensus+, consensus and MAP hypotheses on the Switchboard corpus for two identical systems trained on different amounts of data.

compares the results obtained using the rule-based method with both the MAP baseline and the consensus decoding approach. It can be seen that when the same system is used for producing the training and test data (*Small*), the correction mechanism doubles the gain obtained by the consensus hypothesis over the standard MAP hypothesis. Also, even when the amount of acoustic training data is different for the system that generated the training and test set, we still improve over the consensus hypothesis. This is a useful result when we are not willing to sacrifice acoustic training data for training an error-corrective model.

In addition to improving word error rate, this method has the advantage of producing corrective rules in an easily understood form. For example, the first three rules learned are:

- Choose the second candidate if the first candidate is a short word with a posterior probability less than 0.46 and the second word is "-"(deletion).
- Choose the second candidate if the first word is "A", the second word is "UH" and the difference in posterior probability is less than 0.63.
- Choose the second candidate if the first word is short with a posterior probability less than 0.54 and the second word is long.

These rules can be viewed as diagnostic tools for the speech recognizer, offering the possibility of fixing the cause for these errors instead of correcting the effect. For example, in the second rule involving "A / UH" confusion, the two words share common pronunciations. Therefore, the blame for this error pattern may be attributed to the language model.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we have proposed a new error-corrective mechanism for speech recognition based on transformation-based learning. This is the first approach to utilizing the information existing in confusion networks produced in the consensus decoding framework for discriminating between confusable words. Experiments on the Switchboard corpus show that this new approach results in significant WER reduction when compared with the consensus decoding approach. The acquired knowledge is captured in a small and easily understood set of rules which can serve as diagnostic tools for a speech recognition system. Part of the future work is to add predictive features from the neighbouring context. Our experiments show that in more than 60% of the confusion sets we can predict the correct word with high accuracy. This is an attribute that can be exploited when expanding the current methods to incorporate context-dependent features. The difference in performance between the rule-based methods and probabilistic machine learning methods is also to be investigated.

6. ACKNOWLEDGEMENTS

Many thanks to Stanley Chen, Geoffrey Zweig and Kishore Papineni for valuable comments.

7. REFERENCES

[1] E. Brill. Transformation-based Error-driven Learning and Natural Language: A Case Study in Part of Speech Tagging. In *Computational Linguistics*, 21(4):543-565.
[2] T. Dunning. Accurate Methods for the Statistics of Surprise and Coincidence. In *Computational Linguistics*, 19(1):61-74, 1993.
[3] E. Eide, B. Maison, D. Kanevsky, P. Olsen, S. Chen, L. Mangu, M. Gales, M. Novak and R. Gopinath. Transcription of Broadcast News with a Time Constraint: IBM's 10xRT HUB4 system. In *Proc. ICSLP'00*, Beijing.
[4] V. Goel and W. Byrne. Minimum Bayes-risk Automatic Speech Recognition. In *Computer, Speech and Language*, 14(2):115-135, 2000.
[5] A. Golding and D. Roth. A Winnow-Based Approach to Context-Sensitive Spelling Correction. In *Machine Learning*, San Francisco, 1996.
[6] T. Hain, P.C. Woodland, G. Evermann, D. Povey. The CU-HTK March 2000 Hub5e Transcription Workshop. In *Proc. NIST 2000 Transcription Workshop*, Maryland, 2000.
[7] J. Huang, B. Kingsbury, L. Mangu, M. Padmanabhan, G. Saon and G. Zweig. Recent Improvements in Speech Recognition Performance on Large Conversational Speech. In *Proc. ICSLP'00*, Beijing.
[8] L. Mangu, E. Brill, and A. Stolcke. Finding Consensus in Speech Recognition: Word Error Minimization and Other Applications of Confusion Networks. In *Computer, Speech and Language*, 14(4):373-400, 2000.
[9] L. Mangu and E. Brill. Automatic Rule Acquisition for Spelling Correction. In *Proc. ICML*, Nashville, 1997.
[10] G. Saon, M. Padmanabhan, R. Gopinath and S. Chen. Maximum Likelihood Discriminant Feature Spaces. In *Proc. ICASSP'90*, Turkey.
[11] A. Stolcke, Y. Konig, and M. Weintraub. Explicit Word Error Minimization in N-Best List Rescoring. In *Proc. Eurospeech'97*, pp 163-166, Rhodes.