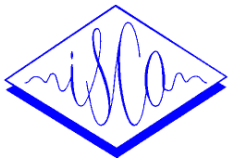

ETRW on
Accessing Information in Spoken Audio
Cambridge, UK
April 19-20, 1999

TASK DEPENDENT LOSS FUNCTIONS IN SPEECH RECOGNITION: APPLICATION TO NAMED ENTITY EXTRACTION

Vaibhava Goel, William Byrne

Center for Language and Speech Processing
Johns Hopkins University
Baltimore, MD 21218
{vgoel, byrne}@mail.clsp.jhu.edu

ABSTRACT

We present a risk-based decoding strategy for the task of Named Entity identification from speech. This approach does not select the most likely utterance produced by an ASR system, which would be the maximum a-posteriori (MAP) strategy, but instead chooses an utterance from an N-best list in an attempt to minimize the Bayes Risk under loss functions derived specifically for the Named Entity task. We describe our experimentation with three risk-based decoders corresponding to the following three performance evaluation criteria: the F-measure, the slot error rate, and the fraction of correctly identified reference slots. An unsupervised optimization is also applied to these decoders. The MAP decoder is used as the baseline for comparison. Our preliminary experiments with these task dependent decoders, using N-best lists of depth 200, show small but encouraging improvements in performance with respect to both manually tagged and machine tagged reference.

1. INTRODUCTION

Identification of Named Entities (NE) in written text or from speech is an important step towards the goals of extracting information, identifying concepts, and ignoring non-information bearing words. NE identification was introduced in the 6^{th} Message Understanding Conference as a component information extraction task, and is referred to as the IE-NE task.

The IE-NE task requires identifying three types of entities: names, temporal expressions, and numeral expressions. The overall task is to identify all instances of these expressions in an input stream of text or speech and assign their constituents to sub-categories. Each entity can be thought of as an object with a fixed number of slots filled by the constituents. An example of a general named entity and its constituent slots is shown in Figure 1. This figure is taken from the user's manual for the message understanding conference scoring software [1].

The identification of Named Entities from speech has recently received much attention. The majority of

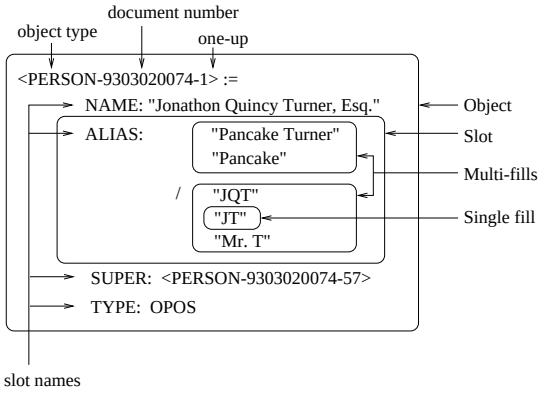

Figure 1: Example of a Named Entity (after [1])

current approaches apply an ASR system to the input speech and then identify the entities in the recognizer output. The performance of IE-NE systems is evaluated at the slot level after aligning the NE tagged recognizer output with manually tagged or machine tagged reference. Various performance evaluation criteria have been used. Some are similar to those used for information retrieval systems, namely precision, recall, and the F-measure, while some others are specific to the IE-NE task, such as the slot error rate (SER) and the fraction of correctly recognized reference slots (FC). For reader's convenience we list the following two criteria [2] [3]

F-measure:

$$F(E', E) = \frac{2\,C(E', E)}{T(E') + T(E)} \tag{1}$$

Slot Error Rate (SER):

$$S(E', E) = \frac{M(E', E) + S_p(E', E) + I(E', E)}{T(E')} \tag{2}$$

where E is the NE tagged recognizer output which is to be compared with the NE tagged reference transcription E'. In the alignment of E with E', $C(E', E)$ is the number of correct slots; $T(E)$ and $T(E')$ are

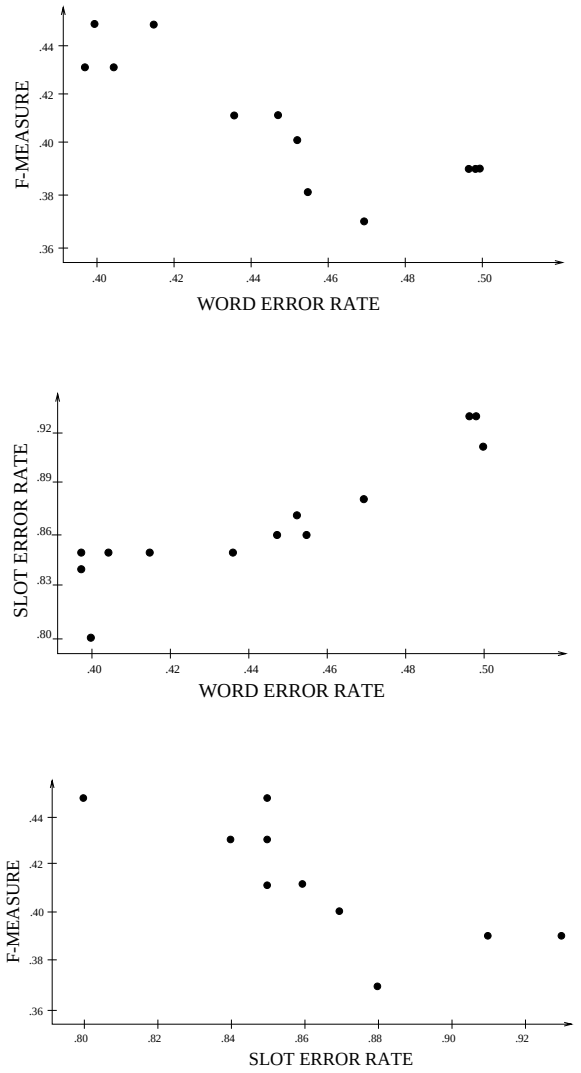

Figure 2: Performance comparison for the thirteen systems submitted for 1998 Hub-5E evaluations (after [7]).

the total number of slots in E and E', respectively; $M(E', E)$ is the number of missing slots from the reference; $S_p(E', E)$ is the number of spurious slots in the recognizer output; and $I(E', E)$ is the number of incorrect or substituted slots.

In recently held LVCSR evaluations [4] thirteen state-of-the-art ASR systems were evaluated as to their suitability for Named Entity identification from speech. The specific NE annotation guidelines used for these evaluations are available by ftp from NIST [5]. The Identifinder(tm) system (described in [6]) developed by BBN was used to identify entities in the output of each system. The IE-NE performance of these systems was summarized by Martin et.al. [7]; we reproduce their analysis here in Figure 2. The top plot in Figure 2 shows the F-measure performance of these systems as a function of their word error rate (WER). Note that the best two systems in F-measure had identical F-measure but were more than 1.5% apart in WER. On the other hand, the two best WER systems had almost identical WER but their F-measures differed by 2%. The WER for systems with similar F-measure varies by as much as 5%. All plots in Figure 2 indicate that, as intuition would suggest, overall performance does follow WER. However, it also appears that performance on different measures is not completely determined by WER.

In this paper we describe a recognition strategy that is matched to the task of IE-NE from speech. It does not select the most likely utterance produced by an ASR system, as is done in the experiments reported above, but instead chooses an utterance from an N-best list in an attempt to minimize the Bayes Risk under loss functions derived specifically for the NE task. Such an approach was first applied to word error rate minimization by Stolcke et.al. [8], and has since been extended to a class of tasks with different task dependent loss functions [9]. It should be noted that in what is presented here we do not vary any of the system parameters such as the acoustic models or the language models; the variation is only in the decoding strategy that selects one of the N-best candidates as the recognizer output. Our goal is to develop a recognition strategy based on task dependent loss functions to better integrate ASR systems into larger language processing and understanding applications. If the ASR systems did work perfectly, this would not be needed. However, we are interested in getting the best IE-NE performance possible from a flawed system; it may be that a perfect system is not necessary.

The next section describes the formulation of our approach and how it is applied to the task of IE-NE from speech. We then present our experiments and preliminary results on 1998 Hub-5E evaluation data. A brief discussion of the results and some speculations are presented at the end.

2. RISK BASED DECODERS FOR IE-NE

We wish to formulate Named Entity identification in speech as a classification task. For this we specify a task dependent, bounded, and real-valued loss function $l(E', E)$ that describes the loss incurred when an acoustic observation A with true Named Entity tagged word sequence E' is classified instead as belonging to the tagged word sequence E. Both E' and E belong to $\mathcal{E}$, the set of all tagged word sequences.

The loss function for a task is closely related to the performance evaluation metric of that task. For our task we chose three evaluation metrics: the F-measure (Equation 1), the slot error rate (Equation 2), and the fraction of correctly identified reference slots. These three have the intuitively obvious loss functions: $(1 - F(E', E))$, $S(E', E)$, and $(1 - Q(E', E))$, where

$$Q(E', E) = \frac{C(E', E)}{T(E')} \tag{3}$$

It is desirable to have a classification rule $\delta(A)$

$$\delta(A) : \mathcal{A} \rightarrow \mathcal{E}, \tag{4}$$

that has the smallest Bayes Risk

$$B(\delta(A)) = E_{P(E,A)}[l(E, \delta(A))]. \tag{5}$$

The Bayes Risk of $\delta(A)$ is the expected loss when $\delta(A)$ is used as the decision rule for data generated under $P(E, A)$; this distribution describes the data that will be encountered in practice. $l(E, \delta(A))$ is a general loss function; in our work here it will be one of the three mentioned above. It is well known that the decision rule that minimizes the Bayes Risk is given by [10]

$$\delta(A) = \underset{E \in \mathcal{E}}{\operatorname{argmin}} \sum_{E' \in \mathcal{E}} l(E', E) P(E'|A). \tag{6}$$

We call this Bayes optimal decoding rule the *risk-based decoder*. Note that for a binary valued loss function, i.e. $l(\cdot, \cdot) = 0$ or 1, the risk-based decoder is the well known maximum a-posteriori probability (MAP) decoder [11] [1]

$$\delta(A) = \underset{E \in \mathcal{E}}{\operatorname{argmax}} P(E|A). \tag{7}$$

In most applications the test set consists of many utterances and A in Equation 6 is the acoustic evidence for the whole test set (i.e. all the test set utterances put together). Similarly, E and E' are word hypotheses tagged with the Named Entities for the entire test set.

Suppose the loss function can be broken down as

$$l(E', E) = \sum_{i=1}^{T} l(E'_i, E_i), \tag{8}$$

where T is the number of test set tokens. Assume also that

$$P(E'_i|A) \approx P(E'_i|A_i) \tag{9}$$

then Equation 6 can be simplified to a decoder for each test set utterance in the following manner

$$\delta(A_i) = \underset{E_i \in \mathcal{E}}{\operatorname{argmin}} \sum_{E'_i \in \mathcal{E}} l(E'_i, E_i) P(E'_i|A_i). \tag{10}$$

Implementing Equation 10 may be infeasible due to large size of set $\mathcal{E}$. The following N-best list rescoring approximation has been proposed earlier [8] [9]

$$\delta(A_i) = \underset{E_i \in \mathcal{E}_i}{\operatorname{argmin}} \sum_{E'_i \in \mathcal{E}'_i} l(E'_i, E_i) P(E'_i|A_i). \tag{11}$$

where $\mathcal{E}_i$ and $\mathcal{E}'_i$ are small, possibly different, subsets of $\mathcal{E}$ containing word sequences with high posterior probabilities given the i^{th} utterance. Equation 11 can be implemented efficiently for a specified loss function if an estimate of the class posterior probabilities can be obtained.

While the loss functions $S(E', E)$ and $Q(E', E)$ are well approximated by Equation 8, $F(E', E)$ can not be explicitly computed as a sum over test set tokens. We propose the following approximation

$$\begin{aligned} F(E', E) &= \frac{2\, C(E', E)}{T(E') + T(E)} \\ &= \frac{\sum_{i=1}^{T} 2\, C(E'_i, E_i)}{\sum_{i=1}^{T} T(E'_i) + \sum_{i=1}^{T} T(E_i)} \\ &\approx \sum_{i=1}^{T} \frac{2\, C(E'_i, E_i)}{T(E'_i) + T(E_i)} \end{aligned} \tag{12}$$

This is a per-utterance approximation to the F-measure.

Like most other systems for IE-NE from speech, we have a two step strategy: first classify A_i into a word sequences W_i and then use a mapping G to identify Named Entities E_i in W_i. It could be G_{manual} in case of manually tagged sentences or G_{tagger} in case of machine tagging, for example the BBN Identifinder(tm). To obtain the class posterior probabilities we used the following approximation:

$$P(E'_i|A_i) \approx P(W'_i|A_i), \tag{13}$$

where W'_i is the word sequence corresponding to E'_i. The quantity $P(W'_i|A_i)$ is estimated from the N-best lists as shown in our earlier work [9]. Note that in the approximation above, there is no $P(E'_i|W'_i)$ term since we are assuming a deterministic relation G between W'_i and E'_i. However our formulation does support probabilistic relation. A probabilistic mapping would be more suitable if sentences were to be classified into more abstract entities, such as concepts, in which case the syntactic and semantic ambiguity would provide the probabilistic component to the assignment.

As presented in one of our earlier papers [9], we introduce a single tuning parameter in the computation of $P(W'_i|A_i)$. This parameter is then optimized in an unsupervised manner under a given loss function. Our implementation of risk-based decoder for IE-NE from speech is summarized by the pseudo-algorithm of Figure 3.

3. EXPERIMENTS AND RESULTS

Our preliminary experiments were performed on conversational speech over telephone. Two corpora - Callhome and Switchboard were used; the test set was the evaluation set for the 1998 Hub-5E fall evaluations. A 200 entry N-best list of hypotheses for each test set utterance was provided to us by BBN. All the experiments reported below were performed on this data. The BBN Identifinder(tm) system [6] was used to tag the N-best lists with Named Entities. The alignment and scoring was performed using the MUC scoring

[1] This is assuming $l(E', E') = 0$, and $l(E', E) = 1\ \forall E' \neq E$.

for i = 1 to T (indices of training set tokens)

1. For A_i, get N-best list of sentences W_i^n and $P(W_i^n, A_i)$.
2. Tag each sentence in the N-best list with the NE-tagger. All N tagged sentences form the set $\mathcal{E}_i^{'}$.
3. Find optimal likelihood tuning parameter by unsupervised optimization.
4. Compute $P(E_i'|A_i)$ for each E_i' in $\mathcal{E}_i^{'}$ incorporating the parameter obtained above.
5. Use Equation 11 to get the desired NE hypothesis.

Figure 3: Pseudo-algorithm for IE-NE in speech

tools developed by MITRE and SAIC, and distributed by NIST. We used the tools included in the IEEVAL0.3 distribution by NIST.

We evaluated our system with respect to two references: manually tagged reference transcriptions and Identifinder(tm) tagged reference. As described above, we investigated three loss functions and their corresponding risk-based decoders: a F-m decoder that optimizes the F-measure (F); a SE decoder that minimizes the slot error rate (S); and a FC decoder that maximizes the fraction of correctly identified reference slots (Q). For F-measure we used the approximation of Equation 12.

Table 1 shows the upper and lower bounds on the performance obtainable from the 200-best lists with respect to the manually tagged reference. The bound on the F-measure performance was obtained by first selecting the sentences corresponding to per-utterance F-measure bounds and then evaluating the resulting set of sentences against the reference under the actual F-measure.

	F (%)	S (%)	Q (%)
Oracle best	63	50	53.1
Oracle worst	12	265	14.4

Table 1: Bounds on the performance from 200-best lists. F is the F-measure, S is the SER, and Q is the fraction of correct reference slots.

The performance of three risk-based decoders with respect to manually tagged and Identifinder(tm) tagged reference is given in Tables 2 and 3, respectively. The baseline performance was obtained using the top, most likely, candidate in each 200-best list.

Although all three decoders can be optimized in an unsupervised manner, we present only optimization results for the slot error rate decoder. Tables 4 and 5 show that this optimization yields a slight performance improvement measures against both manually and Identifinder(tm) tagged reference.

	F (%)	S (%)	Q (%)
Baseline	43	90.5	34.4
F-m decoder	44	86.5	34.4
SE decoder	44	85.9	34.2
FC decoder	28	235.6	46.7

Table 2: Performance of three risk-based decoders on three evaluation metrics with respect to manually tagged reference.

	F (%)	S (%)	Q (%)
Baseline	48	81.8	37.7
F-m decoder	50	77.0	37.8
SE decoder	50	76.7	37.6
FC decoder	30	228.8	49.2

Table 3: Performance of three risk-based decoders on three evaluation metrics with respect to Identifinder(tm) tagged reference.

	F (%)	S (%)	Q (%)
Un-optimized SE decoder	44	85.9	34.2
Optimized SE decoder	44	85.4	33.9

Table 4: Effect of unsupervised optimization on SE decoder. Performance evaluated against manually tagged reference.

	F (%)	S (%)	Q (%)
Un-optimized SE decoder	50	76.7	37.6
Optimized SE decoder	50	76.2	37.4

Table 5: Effect of unsupervised optimization on SE decoder. Performance evaluated against Identifinder(tm) tagged reference.

4. DISCUSSION AND CONCLUSIONS

We have presented a risk-based decoding strategy for identification of Named Entities from speech. This strategy aims at directly optimizing the expected performance of the system under the criterion of interest. We show why the implementation of an exact risk-based strategy may be infeasible and show approximations that implement a per-utterance decoder based on an N-best list rescoring procedure.

Upon comparing the baseline with the oracle numbers it is evident that even in these relatively small N-best lists there is room for substantial increase or decrease in performance. We note also that the overall performance is better when the reference is tagged by the Identifinder(tm) as opposed to when it is tagged manually. This points out a limitation of our system: we use the Identifinder(tm) to tag the recognizer N-best lists (step 2 in the pseudo-algorithm of Figure 3) and hence optimize only for Identifinder(tm). There-

fore, when evaluating the system performance against manually tagged reference, there is a mismatch in optimization and testing criteria.

Looking at the performance of the three decoders we note that even though a rough approximation was used in the F-measure decoder, performance improves by 1 to 2% under the actual F-measure. The SER decoder reduces the slot error rate by about 5%. The FC decoder performs quite well on the evaluation criterion of its interest. However, it produces many spurious slots in an attempt to find as many correct slots as possible, and hence results in a poor SER and F-measure performance. When evaluated with respect to the Identifinder(tm) tagged reference, the F-m and the SE decoders yield a greater increase in performance over the baseline possibly owing to the matched conditions of reference and hypothesis tagging.

We note that our baseline results are not as good as those reported by NIST. This may be due to several factors. The N-Best lists provided to us by BBN were intermediate results and thus did not benefit from all post-processing steps performed to extract the hypotheses submitted to NIST. We note also that the Identifinder(tm) and scoring setup used here may not be identical to those used by NIST.

The techniques presented herein are quite general and apply to a large set of classification problems. As pointed out earlier, it is conceivable to use these ideas to the larger problems of identifying concepts, topics, and stories in speech and text.

5. ACKNOWLEDGMENTS

We would like to thank BBN for providing the Identifinder(tm) system, and we especially Rukmini Iyer and Fred Richardson of BBN for providing us with the 200-best lists for the 1998 Hub-5E evaluation set and the pointers to NIST scoring software.

6. REFERENCES

[1] The message understanding conference scoring software user's manual. Included in the IEEVAL0.3 release by NIST.

[2] C. J. Van Rijsbergen. *Information Retrieval*. 2nd edition, London, Butterworths, 1979.

[3] J. Makhoul, F. Kubala, and R. Schwartz. Performance measures for information extraction. Circulated on the NIST SRU mailing list, June, 1998.

[4] The 9^{th} Hub-5 conversational speech recognition (LVCSR) workshop. MITAGS, Maryland, September 24–25, 1998.

[5] N. Chinchor, P. Robinson, and E. Brown. Hub-4 Named Entity task definition version 4.8. Available at www.nist.gov/speech/hub4_98. 1998.

[6] D. Bikel, S. Miller, R. Schwartz, and R. Weischedel. Nymble: A high-performance learning name-finder. *Fifth Conference on Applied Natural Language Processing*, pp. 194–201, 1997.

[7] A. Martin, J. Fiscus, M. Przybocki, B. Fisher. 1998 Hub-5 workshop: Information extraction. MITAGS, Maryland, September 24–25, 1998.

[8] A. Stolcke, Y. Konig, and M. Weintraub. Explicit word error minimization in N-best list rescoring. *Eurospeech-97*, pp. 163–165, Rhodes, Greece, 1997.

[9] V. Goel, W. Byrne, and S. Khudanpur. LVCSR rescoring with modified loss functions: A decision theoretic perspective. *ICASSP-98*, pp. 425–428, Seattle, WA, 1998.

[10] P. J. Bickel and K. A. Doksum. *Mathematical Statistics: Basic Ideas and Selected topics*. Holden-Day Inc., Oakland, CA, 1977.

[11] L. R. Bahl, F. Jelinek, and R. L. Mercer. A maximum likelihood approach to continuous speech recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 5(2):179–190, 1983.