THE USE OF SENSE IN UNSUPERVISED TRAINING OF ACOUSTIC MODELS FOR ASR SYSTEMS

Rita Singh, Benjamin Lambert, Bhiksha Raj

School of Computer Science, Carnegie Mellon University, Pittsburgh, PA
belamber@cs.cmu.edu, rsingh@cs.cmu.edu, bhiksha@cs.cmu.edu

ABSTRACT

In unsupervised training of ASR systems, no annotated data are assumed to exist. Word-level annotations for training audio are generated iteratively using an ASR system. At each iteration a subset of data judged as having the most reliable transcriptions is selected to train the next set of acoustic models. Data selection however remains a difficult problem, particularly when the error rate of the recognizer providing the initial annotation is very high. In this paper we propose an iterative algorithm that uses a combination of likelihoods and a simple model of sense to select data. We show that the algorithm is effective for unsupervised training of acoustic models, particularly when the initial annotation is highly erroneous. Experiments conducted on Fisher-1 data using initial models from Switchboard, and a vocabulary and LM derived from the Google N-grams, show that performance on a selected held-out test set from Fisher data improves more with iterations relative to likelihood-based data selection.

Index Terms: speech recognition, unsupervised training, sense.

1. INTRODUCTION

In HMM-based automatic speech recognition (ASR) systems, typical acoustic models are a set of Hidden Markov Models (HMMs) that model individual sound units, such as phonemes and triphones. Typically, the HMM states are modeled by Gaussian mixture densities whose parameters are estimated using an expectation maximization (EM) procedure such as the Baum-Welch algorithm. The outcome of this training procedure is dependent on the availability of accurately annotated data (at the word-level) and on good initialization.

Unsupervised training of acoustic models for an ASR system, in the context of this paper, can be thought of as training on untranscribed data. In order to be able to train the system, the data must be automatically transcribed using an initial "bootstrap" model. Training is generally an iterative process. The initial transcriptions from the bootstrap system are used to train acoustic models, which are then used to produce new transcriptions, which are in turn used to update the models and so on. The process is continued until the models become acceptable in quality [1, 2, 3].

The procedure is thus critically dependent on the accuracy of the initial transcriptions obtained from the bootstrap models. When these are highly erroneous, the process can become ineffective. Errors in the initial transcriptions can happen due to many reasons, but the primary reason usually is mismatch between the acoustic and language models used by the bootstrap ASR system and the given speech signals.

For an unsupervised training process to start form a highly erroneous initialization, it stands to reason that a significant component of the unsupervised training process must be the selection of utterances that have been transcribed well enough that they can be used to update the models [2]. This selection can be performed on the basis of a confidence metric [2], or purely based on the likelihood assigned to the transcription by the recognizer [3]. However, when the baseline models used to obtain the initial transcriptions are highly mismatched to the training data, the recognition is typically too poor for either confidence measures or likelihoods (both of which are related) to be useful by themselves.

In this paper, it is our hypothesis that in such situations, it may be possible to use the content of the hypotheses as a criterion for data selection. By the term "content" here, we refer to the "acceptability" of a word sequence: we would like to determine if the hypothesized word sequence would be acceptable in normal spoken contexts, and accept it if so.

Various syntax and semantics-based procedures have been proposed in the literature to judge the validity of recognition hypotheses [4, 5]. These techniques generally attempt to answer the question "could a person have uttered this sequence of words in normal speech", based on some model of language structure or world knowledge [6]. In this paper we take a different approach: instead of asking "could this be spoken by a person", we ask "was this ever spoken in a valid context". If the word pattern found in the recognizer's output has indeed been observed repeatedly in speech or text, we assume that it could be a plausible hypothesis for the recognizer. The primary difference between the two approaches is that while the former attempts to utilize generalizable models of structure and meaning, the latter is strictly observation-based and makes no attempt to generalize. The implicit claim here is that the only way to be sure if a word pattern is acceptable is to have seen it before with sufficient frequency in the language.

For such a frequentist approach to be effective, however, we require a sufficiently large corpus of textual transcriptions of human speech which can be mined to detect the presence and frequency and patterns observed in the recognizer's output. The vast collection of documents on the World-Wide-Web provide a reasonable approximation to such a corpus. By querying the web with the word pattern to be analyzed, the frequency of its occurrence can be determined. A note about conversational speech - the experimental platform for this paper - although the documents on the WWW do not represent informal speech, large portions of it, such as blogs, discussion boards, twitter feeds, etc. do use conversational styles of language. In addition, the web, by nature, is current - it reflects both current usage of language as well as current topics of discussion. The web does indeed also have vast amounts other irrelevant or noisy material as well, but we expect that they are largely unlikely to contain many repetitions of the patterns of the kind that may be output erroneously by a recognizer.

Ideally the word pattern we would search for would be the entire hypothesized word sequence. In the specific situation we address, where the bootstrap recognizer that produces the initial transcription

is very poor to begin with, that would not be effective, however. The recognizer is unlikely to output completely correct sentences that can be validated in entirety by any corpus. Also, the utterances themselves may be segmented out from larger recordings and may not be complete sentences. They may include fragments of sentences, multiple concurrent sentences, or even fragments of multiple concurrent sentences. As a result, instead of looking for the occurrence of entire word sequences, we must restrict ourselves to searching for word sub-patterns that occur in the recognizer hypotheses.

We note that a specific instance of this approach has been highly popular in the speech recognition community. Word N-grams are short word-sequence sub-patterns. Every modern large vocabulary ASR system effectively uses the relative frequency of N-word sequences, as determined from a large corpus of training text, to provide cues for recognition. In our work however, we propose to search for a more generalized set of patterns, including sequential patterns.

Another annoying problem that prevents us from using the web as-is in the manner proposed above, is logistical. It is currently practically infeasible for us to search the web and parse returned documents to analyze every sentence hypothesized by the recognizer. We therefore use the Google N-gram data [7] which reports counts of word sequences up to 5 words long from a snapshot of the web circa 2006 as a surrogate for the web. While this falls short of our intended goal of searching for all patterns in hypothesized word sequences, since it restricts us to searching for word patterns that span no more than five words, it is a reasonable approximation for the concepts explored in this paper.

Briefly, in this paper we explore the use of a content-based sensicality measurement of automatic transcriptions for data selection in unsupervised training. We evaluate the sensicality score of a hypothesized word sequence using an empirically derived function of the normalized frequency of occurrences of every word subsequence up to five words long. The likelihood of word sequences, although not entirely reliable by itself, also provides an important cue and is also incorporated into the score. Utterances are selected if both the sensicality score, and the acoustic likelihoods of their current hypothesized transcriptions exceed prespecified thresholds. The entire unsupervised training procedure is thus iterative. At each iteration including the first, all hypotheses whose sensicality score exceeds a threshold are selected for the next iteration. As we demonstrate in the experimental section, selection based on sensicality score alone, and on sensicality and likehihoods, results in better trained models, as measured by recognition performance on a held-out set, than selection based on likelihood alone, or by selecting all hypotheses. Interestingly, we also observe that the number of hypotheses we are able to select also increases with iterations, showing that procedure is effectively a hill-climbing procedure over sensicality.

The rest of this paper is organized as follows. In Section 2 we explain how we define and compute our sensicality score and how it is used to select training data. In Section 3 we outline the entire iterative unsupervised learning procedure that follows. In Section 4 we describe our experiments and in Section 5 we present our conclusions.

2. A CONTENT-BASED SELECTION METRIC

Our objective is to design a mechanism to evaluate the acceptability of a hypothesized word transcription based directly on its content, rather than any correlate such as likelihood or confidence measures that have been derived primarily from acoustic measurements.

Although humans appear to be very good at such judgment [8], content-based computational models for measuring the acceptability of recognition hypotheses remain elusive. A vast array of literature currently exists on characterizing both the structure e.g. [5] and semantic coherence [9] of word sequences, yet, to our frustration, neither of them provides a sufficiently generic model to judge the acceptability of word sequences that are obtained as speech recognition hypotheses. The following trivial examples illustrate the problem. The sentence "two times two is five" is perfectly structured from the perspective of following the dictates of English grammar. But if a recognizer were to hypothesize it one would be inclined to believe that the hypothesis is incorrect and that the "five" is actually a misrecognition of the word "four". That is because the hypothesis violates our world knowledge - it is not semantically coherent. On the other hand, if a recognizer were to hypothesize "colorless green ideas sleep furiously", one would be inclined to accept it although the sentence famously violates world knowledge in every possible way. A word sequence with such detailed structure is unlikely to have been hypothesized by happenstance; if it is hypothesized, a person is most likely to have spoken it.

We use an alternate approach to determine if a word sequence may be accepted or not. We claim that one near-certain criterion to determine if a word sequence is valid is to determine if the word sequence has been spoken (or written) earlier. Minimally, the word patterns in the word sequence must have been observed. In order to make this determination, we require a very large "search" corpus of text, in our case preferably of the style of spoken speech, which can be searched for the word patterns. The world wide web provides us with this corpus. It comprises an ever expanding set of documents, many of which are in conversational style.

In practice, however, it is currently impractical for us to search the web repeatedly to evaluate every hypothesized word sequence. Instead, we use the Google N-gram data [7] as a proxy for the web. The Google N-gram data include counts of all unigrams, bigrams, trigrams, quadgrams and quingrams that occur a minimum of 40 times in a snapshot of the web circa 2006. The N-grams were derived from a corpus of over one trillion words. Thus, although it does impose on us the restriction that the largest word patterns we can look for span five words, we can be reasonably confident about the robustness of any measure derived from them.

The actual procedure for content-based selection is as follows: given a word sequence $w_1 w_2 \cdots w_K$ we form L-word-long patterns of the kind $w_i w_{i+1} \cdots w_{i+L-1}$ where L can be any number between 2 and 5 and i can be any number between 1 and $K - L$. Let $P(i, L)$ be the pattern representing the word sequence beginning at the i-th word and extending for L words. If the word sequence were perfectly formed, we would expect all of these patterns to be present in the search corpus. However, we must take into consideration that i) the recognizer makes many errors. We are willing to accept hypotheses that have some errors, provided they are relatively few in number, ii) the search corpus is not comprehensive in spite of its size. Note that to account for (i) adequately, we must in principle also account for insertions and deletions in the hypothesis, which effectively transforms the patterns into regular expressions; however we do not do so in this paper.

We also consider the fact that longer patterns are more important than shorter patterns. Thus, the detection of a longer pattern is more important than that of a shorter one. Finally, we must also consider that longer hypothesized word sequences will naturally have more patterns which could potentially be discovered in the search corpus.

Based on the above criteria, we define an empirically derived

"sensicality score" for the word sequence as

$$Sensicality(w_1 w_2 \cdots w_K) = \frac{1}{K} \sum_{L=1}^{5} \sum_{i=1}^{K-L} LCcount(P(i, L))$$

Where $count(P(i, L))$ is the count of the number of times the pattern $P(i, L)$ has been observed in the WWW as given by the Google N-gram counts. C is a scaling constant. We select all utterances for which this sensicality score is greater than a threshold.

The sensicality measure given above does not account for the actual acoustic log likelihood assigned to the word sequence in the process of recognition. Although log likelihood is not a reliable measure to detect the correctness of a hypothesis, it does provide us with a rough indication of the degree to which the acoustics matched the word sequence hypothesized. This is required to ensure that the word sequence for the selected utterance was generated from acoustic match and not by chance. We therefore also incorporate the condition that the per-frame log likelihood for the utterance must exceed a threshold. The overall condition for selecting an utterance U that has been automatically transcribed as $w_1 w_2 \cdots w_K$ is thus given by the Boolean equation:

$$Select(U, w_1 w_2 \cdots w_K) = (L_p > \theta_L) \&\& (sensicality(w_1 w_2 \cdots w_K) > \theta_S)$$

where L_p is the *per-frame* log likelihood, and $_L$ and θ_S are the thresholds on the log likelihoods and sensicality. We note that the suggested procedure is essentially one of selection. Since selection is binary, it cannot be used for rescoring recognition hypotheses or to provide confidence measurements. In essence, it is specific to the problem of data selection, as in the case of unsupervised training.

Here are examples of sentences selected with a high sensicality score. Note that these sentences are not necessarily complete or meaningful; however they clearly are acceptable as parts of a human speech

score	sentence
7.99	and you would want to go with them so
7.67	but a lot of what you have to be working with that
7.58	i mean this is having a sense of what you doing

Here are examples of low-scoring sentences that were rejected.

score	sentence
0.75	everything's just think clan or
0.75	yeah she's she's been a intact first area other
0.74	yeah not o'riley don't really get toxin videotapes and those are and

Ideally, utterances whose hypothesized transcriptions are discovered, either in full or in large part, in the corpus, will be retained and the rest rejected. It is expected that initially, when the models are poor, a large fraction of the utterances are likely to be discarded. We expect the number of selected utterances to increase as the models improve.

3. UNSUPERVISED TRAINING AS AN ITERATIVE HILL-CLIMBING PROCEDURE

The overall unsupervised training procedure is an iterative process. In the first iteration a bootstrap acoustic and language model are used to produce initial transcriptions for the entire untranscribed training data. The selection procedure described in Section 2 is then used to select utterances from the training set. The selected utterances are used to train an acoustic model. The newly trained acoustic model is then used to recognize the entire training data (including the utterances that were not selected for training in the first round). The data selection procedure from Section 2 is used once again to select utterances and the entire process is repeated.

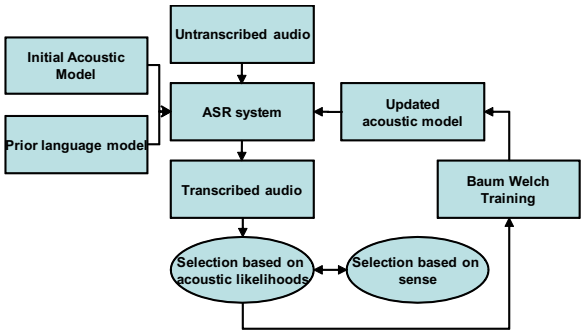

Fig. 1. Flowchart of iterative training procedure.

The actual training procedure employed within each iteration is the Baum-Welch training procedure which is itself iterative. Using the conventional training procedure for acoustic models, we train context-independent models with the selected utterances (which require several iteration of Baum-Welch estimation), followed by context-dependent models with no parameter sharing. These "untied" context dependent models are used to build decision trees for state tying, following which the final tied-state models are trained. It is the tied-state models that are used to obtain the transcriptions for the next iteration of the algorithm. The overall iterative algorithm is shown as a flow chart in Figure 1.

4. EXPERIMENTAL RESULTS

Our objective is to design a content-based data selection procedure for unsupervised training in high initial transcription error situations. We chose the Fisher Phase I corpus available from LDC (Catalog No. LDC2004S13) as our training set for the experiment. The data include 5,850 two-channel audio files, each one containing a full conversation of up to 10 minutes. The channels were separated and segmented into short segments. We used a total of 111157 speech segments, representing nearly the entire data, minus our held out test set, as the training data. While the transcriptions for this corpus have been provided by LDC, we assumed that no transcriptions were available. We identified a set of 10,000 segments from the same data as our held-out designated test set. For the test set, we used the reference transcriptions provided by LDC for computing Word Error Rates (WERs) of decoded test hypotheses.

We used the CMU Sphinx-3 ASR system to perform all our experiments. Acoustic models for all iterations, excluding the initial one, consisted of triphone HMMs for a total of 162900 triphones with 44 base phones. Each HMM had a 3-state left-to-right Bakis topology with no skips permitted across states. State output distributions were Gaussian mixture densities with 16 Gaussians/mixture.

Since the the Fisher corpus consists of spontaneous telephone bandwidth speech, we chose acoustic models trained from the switchboard corpus (LDC Catalog 97S62), to produce initial transcriptions of the training data This bootstrap model set consisted of triphone HMMs with 3-state left-to-right Bakis toplogy HMMs. The model had 5000 tied states, each modelled by a mixture of 16 Gaussians, trained using a maximum likelihood Baum Welch procedure. On the switchboard test corpus, this set of acoustic models, with a

Iter	WER	Accuracy	No. of utts selected
0 (initial)	88.5	16.1	111k
1	87.1	18.5	35k
2	86.4	18.7	37k
3	84.4	21.4	45k
4	82.4	22.8	49k
5	79.2	31.0	70k
6	77.1	31.0	72k
7	75.3	31.0	74k

Table 1. Sensicality+likelihood based data selection

Iter	WER	Accuracy	No. of utts selected
0 (initial)	88.5	16.1	111157
1	79.1	28.7	55081
2	79.5	28.1	61270
3	80.0	27.4	65934
4	79.3	29.5	71293
5	79.5	30.5	73233

Table 2. Likelihood based data selection

standard CMU internal SWB language model, gives an error rate of 45%. However, these acoustic models are highly mismatched with the Fisher data and in combination with the LM described below result in very poor recognition accuracies on the Fisher training set. This poor accuracy was not incidental, but by design, since our intention is to simulate a situation where the initial transcriptions obtained in the unsupervised training process are *highly* erroneous.

The experiments also require a language model. For the selection of the language model, we assumed to have no knowledge of the vocabulary of the training corpus, and merely used the fact that it was conversational speech. We therefore used a language model constructed for conversational speech to recognize the standard Switchboard corpus in-house at Carnegie Mellon university in the year 2001. The same LM was used in all experiments.

We used the combination of likelihood and sensicality proposed in Section 2. In the first iteration models were trained with the initial automatically obtained transcriptions. Thereafter, at each iteration the models from the previous iteration were used to recognize the *entire* training set. Data were selected from the recognized utterances based on the combined sensicality+likelihood metric proposed in Section 2. In order to be selected, an utterance had to have a per-frame likleihood greater than a threshold value of 0. Utterances that exceeded this threshold were further required to have a sensicality score greater than a sensicality threshold. We used an empirically derived sensicality threshold of 1.3 for selection. Utterances that were selected using this procedure were used to train the next iteration of acoustic models.

Table 1 shows the progression of the models with iterations. The first column of the table shows the word error rate obtained on the held-out test set. The WER computed here accounts for both insertions and deletions. Since, for unsupervised training the *recall* of the recognizer is important, the table also shows the recall (accuracy) in the second column. The final column shows the number of utterances selected by our criterion.

What would we get if we used only likelihood-based selection. Table 2 shows what happens. For the results in this table, we used only the likelihood criterion to select data (using a likelihood threshold of 0), based on the same recognition outputs used for Table 1. Each row of Table 1 corresponds to the same iteration reported in Table 1. We immediately note that after the first iteration, the recognition performance of the system ceases to improve although the number of utterances that have a likelihood greater than the threshold and are therefore selected continues to increase. In the early stages of the training, the sensicality-based selection actually lags behind likelihood-based selection. But as the iterations progress, whereas the likelihood-based technique saturates quickly, the sensicality-based selection continues to improve, presumably because of the injection of external information from the web. In Table 1 the number of selected sentences increases with iterations, suggesting that the procedure is a hill-climbing one in sensicality.

5. CONCLUSIONS

We have proposed a "sensicality" based measure for data selection in unsupervised training. The sensicality measure measures not so much sense as whether the word patterns in a recognition hypothesis have been observed in real speech or text earlier. To make this determination, the Google Ngrams, a derivative of the WWW is used.

The experiments we run show that selecting data in this manner results in significantly greater improvement in recognition performance with iterations of unsupervised training than likelihood-based data selection. In other results not shown here for lack of space, both likelihood and sensicality based selection is also superior to *not* selecting data at all and simply using all data. Subjectively, we also observe that the sensicality measure correlates well with human judgement of meaningfulness – sentences that score highly on the sensicality measure generally appear more semantically coherent than those that did not. Once again, examples have not been provided for lack of space. In general, we believe the use of such extra-acoustic information for unsupervised training to be a rich pasture.

Most of all, a take-away lesson we got from this exercise was the viability of the WWW as a terrific resource for NLP tasks, in a manner not hitherto possible. The sheer size of it makes searches of whole-sentence and partial-sentence patterns possible among other things. Although we have only used a highly reduced proxy of it, we nevertheless observe its usefulness immediately. This resource opens a number of options that we plan to explore in future.

6. REFERENCES

[1] F. Wessel and H. Ney., "Unsupervised training of acoustic models for large vocabulary continuous speech recognition", Proc. ASRU 2001.

[2] J. Ma, R. Schartz., "Unspervised training on a large amount of Arabic news broadcast data", Proc. ICASSP 2007.

[3] B. Xiang, Y. Deng, and Y. Gao., "Unsupervised training for Farsi-English speech-to-speech systems", Proc. ICASSP 2006.

[4] R. Rosenfeld., "Adaptive statistical language modeling: a Maximum Entropy approach," Ph.D thesis, CMU 1996.

[5] C. Chelba and F. Jelinek, "Structured language modeling,", Computer Speech and Language, 2000.

[6] D. Bianchi, "Ontology based automatic speech recognition and generation for human-agent interaction," WETICE 2004.

[7] T. Brants and A. Franz., "Wet 1T 5-gram Version 1", Linguistic Data Consortium, Philadelphia 2006.

[8] E. Brill, R. Florian, J. C. Henderson and L. Mangu, "Beyond N-grams: Can linguistic sophistication improve language modelling?", COLING 1998.

[9] I. Gurevych, R. Malaka, R. Porzel., "Semantic Coherence Scoring Using and Ontology," HLT-NAACL 2003.