N-BEST LIST RERANKING USING HIGHER LEVEL PHONETIC, LEXICAL, SYNTACTIC AND SEMANTIC KNOWLEDGE SOURCES

Mithun Balakrishna and Dan Moldovan

The University of Texas at Dallas
Richardson, Texas 75080 USA
{mithun,moldovan}@hlt.utdallas.edu

Ellis K. Cave

Intervoice Inc.
Dallas, Texas 75252 USA
Skip.Cave@intervoice.com

ABSTRACT

This paper presents a novel methodology to improve Large Vocabulary Continuous Speech Recognizer (LVCSR) hypotheses using additional phonetic, lexical, syntactic and semantic knowledge. Such additional higher level knowledge sources are unavailable during the LVCSR decoding due to the various constraints placed on the successful deployment of such information sources. This paper will focus on the extraction of WER improvements from the LVCSR n-best list using the additional higher level knowledge sources as the nucleus of a reranking mechanism. We shall illustrate the improvements obtained for the conversational speech transcription task and also for the directed dialog speech utterance transcription task in a grammar tuning application.

1. INTRODUCTION

Current generation of Large Vocabulary Continuous Speech Recognition (LVCSR) systems rely on a HMM based acoustic model and a n-gram based language model (LM) to perform speech-to-text conversion with reasonable accuracy. Several LVCSR systems, with accuracy constraint, perform multiple recognition passes on each speech utterance to get the best performance with respect to Word Error Rate (WER). Even in such multi-pass LVCSR systems, a significant amount of untapped WER improvements remain hidden inside the LVCSR n-best lists and word-lattices. We use SONIC [1], a LVCSR from the University of Colorado at Boulder, to find the extent of WER improvements that could be extracted from the n-best lists and word lattices for the NIST HUB5 2000 evaluation set (40 CallHome and 40 Switchboard-1 conversation sides).

We train SONIC for the telephone transcription task using 160 CallHome conversation sides and 4826 Switchboard-1 Release 2 conversation sides (without the speech files from HUB5 2000 set). We use the SRI HUB5 2000 model as the tri-gram back-off LM. SONIC performs a 3-pass recognition (Gender-dependent acoustic models, VTLN + SAT acoustic models + MLLR, MLLR(2) adaptation) to produce the baseline hypotheses, n-best lists and word lattices.

Figure 1 presents the Oracle WER values possible at various depths in a 200-best list produced by the 3-pass SONIC LVCSR for the HUB-5 set. The oracle uses manual speech transcriptions as reference to propose the best hypothesis at various depths in the n-best list. The experimental analysis for Switchboard (CallHome) shows that the choice of the best hypothesis from the 200-best list gives a 10.12% (13.59%) absolute WER reduction over the 3-pass recognition baseline, while the word lattice gives a 11.09% (14.66%) absolute WER reduction. The Oracle captures majority of the most accurate Switchboard (CallHome) hypotheses within 80 to 90 n-best depth for each utterance. These experimental results prove that substantial improvements can be gained by applying a strong post-processing mechanism like reranking, even at a small n-best depth.

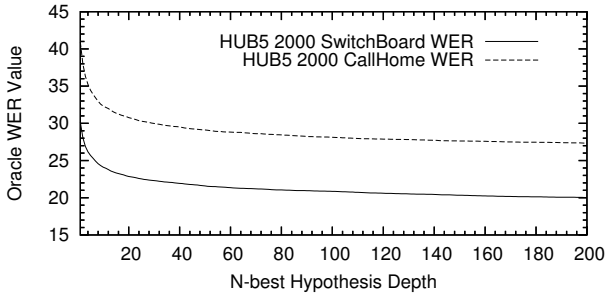

Fig. 1. Oracle WER at various n-best depths for HUB5 2000 set.

2. N-BEST LIST RERANKING

There have been several attempts at combining various knowledge sources, unavailable to the original ASR decoder, into a post-ASR mechanism to reduce the original WER. Reference [2] proposes two semantic LMs: a semantic concept based model using long span semantic units to capture an utterance's meaning sequences and a semantic structured model which uses semantic parsers to extract information from an utterance. Similar to the lexical/syntactic filtering mechanisms and semantic transformation based LM for the Question-Answering domain [3], [4] propose a syllable based lexical model and a domain based lexico-semantic oriented approach to process lexical/semantic errors. In this paper, we avoid crafting domain specific linguistic or lexical rules [3, 4] or, model small domain dependent statistical knowledge sources [2] due to their sub-optimal performances in the conversational speech recognition (CSR) task.

Reference [5] focuses on combining n-gram based local dependencies, topic based and syntactic based [6] long distance dependencies into a maximum entropy LM. Reference [7] presents an almost-parsing LM within the Constraint Dependency Framework for CSR, achieving performances competitive with state-of-the-art parser LMs [6, 8]. The LM in [7] tightly integrates multiple knowledge sources but relies mainly on lexical features along with syntactic constraints and slightly on semantic constraints. Given the CSR domain independence, we model multiple knowledge sources [9] into a n-best reranking process (ignoring additional small improvements in word-lattices) and focus on more extensive phonetic and semantic reranking features. Figure 2 captures our proposed architecture of deploy-

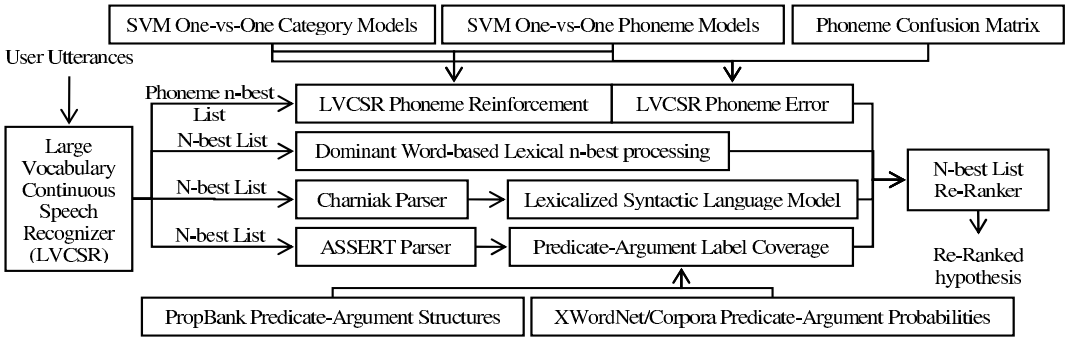

Fig. 2. N-best list reranking process using the phonetic, lexical, syntactic and semantic features.

ing phonetic, lexical, syntactic and semantic knowledge sources. We hope to reduce LVCSR WER by working these sources in tandem, complementing each other.

For every n-best hypothesis, we compute the confidence score in the following manner: if $Ph = ph_1, ph_2,\ldots,ph_m$ is the phoneme sequence corresponding to the word sequence $W_h = w_1, w_2,\ldots,w_k$, predicted by the LVCSR for the acoustic frames $A = a_1, a_2,\ldots,a_m$ of a single utterance, then

$$\begin{aligned} Score\,(W_h) &= P\,(W \wedge Ph|A) = \frac{P\,(W \wedge Ph \wedge A)}{P\,(A)} \\ &= \frac{P\,(Ph \wedge A)}{P\,(A)} * \frac{P\,(W \wedge Ph \wedge A)}{P\,(Ph \wedge A)} \\ &= P\,(Ph|A) * \frac{P\,(Ph \wedge A|W) * P\,(W)}{P\,(Ph \wedge A)} \\ &= P\,(Ph|A) * \frac{P\,(Ph \wedge A|W)}{P\,(Ph \wedge A)} * P\,(W) \quad (1) \end{aligned}$$

The following sections will present a detailed explanation of the knowledge resources modeled for computing each constituent of (1).

2.1. Phonetic Features

We assign phonetic scores to each n-best list hypothesis as a measure of the LVCSR phoneme classification accuracy compared to an additional phoneme classifier. We use Support Vector Machine (SVM) as the additional phoneme classifier due to its superior performance in classification tasks including phoneme recognition [10]. The first objective of the phonetic score is to derive the SVM posterior probability of the acoustic frames fitting into the LVCSR phoneme sequence i.e. assign SVM supportive probability score to the phoneme sequence produced by the LVCSR. The second objective is to categorize the acoustic frames into a sequence of phonemes independent of the original LVCSR phoneme classification. i.e. the best SVM phoneme sequence is used to assign a SVM supportive value based on the match/mismatch with the LVCSR phoneme sequence. We use only a subset of the acoustic frames (a central band of frames) representing each phoneme in W_h for the phonetic score computation due the phoneme context independence assumption in the SVM models.

Training One-Vs-All SVM phoneme models is impractical, cumbersome and has a high error rate. Distributing the task to produce One-Vs-One binary SVM models needs 1225 trained models for a set of 50 phonemes. Hence, we group the phonemes into 13 categories based on their acoustic properties [9] and thus reduce the number of SVM models to 169 (78 One-Vs-One category models and, 91 One-Vs-One models for phoneme pairs in each category).

2.1.1. SVM Phoneme Class Posterior Probability

We use the term $P(Ph|A)$ in (1) to derive the supportive SVM posterior probability of the acoustic frames A fitting into the LVCSR phoneme sequence Ph. Phoneme boundary information in the n-best list is used to determine the LVCSR phoneme sequence for frames A in W_h. $P(Ph|A)$ is modified due to the acoustic property-based hierarchical SVM classification: If $Cat(ph)$ is the function which assigns a category to a phoneme ph, then

$$\begin{aligned} P\,(Ph|A) &= P\,(Ph \wedge Cat\,(Ph)\,|A) \\ &= \frac{P\,(Cat\,(Ph) \wedge A)}{P\,(A)} * \frac{P\,(Ph \wedge Cat\,(Ph) \wedge A)}{P\,(Cat\,(Ph) \wedge A)} \\ &= P\,(Cat\,(Ph)\,|A) * P\,(Ph|Cat\,(Ph) \wedge A) \quad (2) \end{aligned}$$

$P(Cat(Ph)|A)$ in (2) represents the SVM class posterior probability of the acoustic frames belonging to the original LVCSR phoneme category sequence, while $P(Ph|Cat(Ph) \wedge A)$ represents the SVM class posterior probability of the acoustic frames belonging to the original LVCSR phoneme sequence given the phoneme categories.

2.1.2. LVCSR-SVM Phoneme Classification Accuracy Probability

Let $Ph^s = ph_1^s, ph_2^s,\ldots,ph_m^s$ be the best SVM phoneme sequence for A but not containing the phonemes from W's phoneme sequence i.e. each acoustic frame a_i is labeled as a phoneme ph_i or ph_i^s.

$$\begin{aligned} \frac{P\,(Ph \wedge A|W)}{P\,(Ph \wedge A)} &\approx \frac{P\,(Ph \wedge A|\neg Ph^s)}{P\,(Ph \wedge A)} \{\text{Using the defn of} Ph^s\} \\ &= \frac{P\,(Ph \wedge A) - (P\,(Ph \wedge A|Ph^s) * P\,(Ph^s))}{P\,(\neg Ph^s) * P\,(Ph \wedge A)} \\ &= \frac{1}{P\,(\neg Ph^s)} - \frac{P\,(Ph \wedge A|Ph^s) * P\,(Ph^s)}{P\,(Ph \wedge A) * P\,(\neg Ph^s)} \\ &= \frac{1}{P\,(\neg Ph^s)} - \frac{P\,(Ph \wedge A \wedge Ph^s)}{P\,(Ph \wedge A) * P\,(\neg Ph^s)} \\ &= \frac{1}{P\,(\neg Ph^s)} - \frac{P\,(Ph \wedge Ph^s|A)}{P\,(Ph|A) * P\,(\neg Ph^s)} \\ &= \frac{1}{P\,(\neg Ph^s)} - \frac{P\,(Ph^s|A \wedge Ph) * P\,(Ph|A)}{P\,(Ph|A) * P\,(\neg Ph^s)} \end{aligned}$$

$$= \frac{1 - P\left(Ph^s|A \wedge Ph\right)}{P\left(\neg Ph^s\right)} = \frac{1 - P\left(Ph^s|A \wedge Ph\right)}{1 - P\left(Ph^s\right)}$$
$$\approx \frac{1 - P\left(Ph^s|A \wedge Ph\right)}{1 - P\left(Ph^s|Ph\right)} \quad (3)$$

The phonetic score (3) prefers a hypothesis W_h which matches with the best SVM phoneme sequence Ph^{svm} (Ph^{svm} is different from Ph^s). $P\left(Ph^s|A \wedge Ph\right)$ is the probability of acoustic frames A being categorized into Ph^s, given the restriction imposed on SVM regarding the selection of Ph^s, by the LVCSR sequence Ph in W_h (from Ph^s defn). We compute $P\left(Ph^s|Ph\right)$ using a matrix, containing the probabilities of a frame a_i mapping to ph_i^s given the LVCSR choice ph_i, collected over a large training set. The probability value $P\left(Ph^s|A \wedge Ph\right)$ is low if ph_i matches with the best SVM phoneme ph_i^{svm} for a_i, since this would imply that ph_i^s is the SVM's second best choice, and hence has a lower SVM posterior probability value than ph_i^{svm}. Intuitively, $P\left(Ph^s|Ph\right)$ has a high value due to the increased number of matches between ph_i^{svm} and ph_i, and hence increased number of conflicts between ph_i^s and ph_i. Thus, the score (3) assigned to W_h will be high. Similarly, (3) is low when ph_i does not match with ph_i^{svm} for a_i. In (3), we compute $P\left(Ph^s|A \wedge Ph\right)$ using the category and phoneme SVM models just like in (2).

2.2. Lexical Features

Reference [11] approximates the posterior probabilities and then computes the expected WER using the posterior probability distribution. Agreeing with the views in [11] on extracting certain suppressed n-gram word combinations, we propose a simple mechanism (lexical features have a small performance gain with huge computational need [9]) to capture dominant word occurrences using the n-best list word boundary information (avoids string alignment) and score each hypothesis based on the presence of these dominant words.

$$Score_{Lx}\left(W_h\right) = P_{LM}\left(W_h\right) * \prod_{i=1}^{k} \frac{P\left(w_h^i \mid B\left(w_h^i\right)\right)}{P\left(W \mid B\left(w_h^i\right)\right)} \quad (4)$$

$P_{LM}(W_h)$ in (4) is the LVCSR trigram score associated with the n-best list hypothesis containing k words. In the second part of this score, we compute the unigram probabilities of the words in each hypothesis, given their word time boundaries, using the set of words occurring in the n-best list as a reference i.e. probability of the word w_h^i occurring in a particular hypothesis time-frame, normalized by all the n-best list words W occurring in that particular hypothesis time-frame in all the n-best hypotheses. Function $B\left(w_h^i\right)$ returns the time boundary information for the word w_h^i usually with a buffer of a small constant number of time frames on either side of the word.

2.3. Syntactic Features

The replacement/interpolation of syntactic models [5, 6, 7, 8, 12] with n-gram LMs, for n-best list reranking, word lattice processing, or directly into the LVCSR, has seen a large amount of literature lately. Reference [8] presents a two-stage parsing LM. They generate a set of high probability candidate parses using a PCFG parser and then rescore them using a lexicalized syntactic LM [12].

Since immediate-head parsers [12] are the most accurate among other statistical parsers, the idea of using a syntactic LM based on immediate head parsing seems too attractive to be ignored. This is ideal for our work since the n-best reranking does not impose a left-to-right processing constraint. We train a lexicalized syntactic LM for conversational speech using the immediate-head parser and generate the syntactic score for each n-best hypothesis. Hence, the syntactic score (5) is given by the probability of the best parse (π), for the hypothesis W_h, generated by the lexicalized syntactic model.

$$Score_{Syn}\left(W_h\right) = P\left(\pi, W_h\right) \quad (5)$$

2.4. Semantic Features

Primarily, the use of semantic knowledge in ASR LMs has been limited to semantic representations of an application domain, depending on the generation/availability of such domain dependent semantic knowledge [2, 3, 4]. In this paper, we extract semantic knowledge from the basic semantic propositions of a language and use it for the n-best reranking. PropBank [13] is a 300K word corpus which contains predicate argument relations and labeling for the verbs from the WSJ part of Penn TreeBank. For our n-best reranking, we are interested in the semantic argument assignments defined in PropBank for around 3500 verbs. For these verbs, PropBank assigns a set of core arguments and additional adjunctive arguments in some cases.

[14] present a semantic parsing technique using SVM, to identify and label semantic arguments for each predicate in a sentence. The task of the semantic parser is to identify the constituents of a sentence that represent the semantic arguments of a given predicate and assign appropriate argument labels to them. We use the ASSERT semantic parser [14] to extract statistical semantic knowledge, based on the predicate-argument relations defined in PropBank. We ran the semantic parser on a 3 million sentence text corpus (Switchboard & CallHome transcriptions excluding HUB5 evaluation test set and, LA Times Corpus) and, for each identified predicate, count the various argument labels covered, using the PropBank list of arguments for verbs as a reference. We produce a statistical knowledge source containing 6.9 million identified predicate argument structure.

EXtended WordNet2.0-1 [15] enhances WordNet2.0. It contains enhancements of the Wordnet glosses, subject and direct object labeling for predicates in the glosses, syntactically parsed glosses and links with other glosses that describe related concepts. The previously extracted semantic knowledge is augmented with additional predicate argument coverage information (117k predicates) from eXtended WordNet2.0-1. The extracted semantic knowledge is used to evaluate the semantic coherence of the n-best hypotheses.

$$\begin{aligned} Score_{Sem}\left(W_h\right) &= \sum_{i=1}^{r} P\left(Coverage\left(W_h, v_i\right) \mid v_i\right) \\ &= \sum_{i=1}^{r} \frac{P\left(Coverage\left(W_h, v_i\right) \wedge v_i\right)}{P\left(v_i\right)} \\ &= \sum_{i=1}^{r} \frac{\frac{\sum_{j=1}^{p} Cnt(FillSlot(j, W_h, v_i) \wedge v_i)}{\sum_{j=1}^{q} Cnt(FillSlot(j, v_i) \wedge v_i)}}{\frac{Cnt(v_i)}{\sum_{j=1}^{t} Cnt\left(v_j\right)}} \end{aligned} \quad (6)$$

We run each n-best hypothesis through the semantic parser and find the arguments for all the predicates identified in the hypothesis. In (6), $v_1, v_2, \ldots, v_r$ are the predicates in the LVCSR hypothesis W_h, p is the number of arguments for predicate v_i in hypothesis W_h, q is the total number of arguments possible for predicate v_i in PropBank and t is the total number of predicates identified in PropBank. (6) provides a semantic score based on the various argument labels corresponding to all the predicates v_i identified in the hypothesis W_h. $FillSlot\left(j, W_h, v_i\right)$ lists all the occurrences of the predicate v_i filling j arguments labels in W_h, while $FillSlot\left(j, v_i\right)$ lists all the occurrences of the predicate v_i filling j arguments labels in the extracted semantic knowledge base.

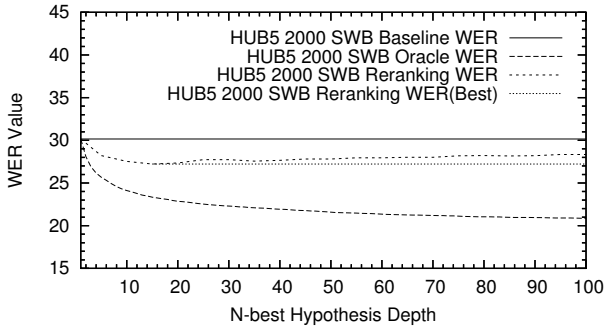

Fig. 3. Various HUB5 2000 Switchboard evaluation WER results.

3. RESULTS AND EXPERIMENTAL SETTINGS

The reranking score assigned to an LVCSR n-best hypothesis is a simple linear weighed combination of the individual scores from each participating knowledge source. $(w_1, w_2), w_3, w_4, w_5$ represent the weights assigned to the phonetic (two scores), lexical, syntactic and semantic features respectively.

The baseline configuration of SONIC remains the same as the one detailed in section 1. The SVM models are trained using the 39 PMVDR [1] feature set while the SONIC models are trained using the conventional 39 MFCC feature set. We create the 169 One-Vs-One SVM models using 64K positive and 64K negative examples from the SONIC training speech files. We described the experimental settings for the phonetic, lexical, syntactic and semantic features in the previous sections. Figure 3 presents the reranking results at various n-best depths (varied by multiples of 5) for 40 HUB5 2000 Switchboard conversation sides using the best knowledge weights combination of $(w_1 = 16, w_2 = 17, w_3 = 14, w_4 = 18, w_5 = 35)$. We find the best set of weights by running WER testing trials on 40 HUB5 2000 CallHome conversation sides as a development corpus.

The proposed reranking mechanism achieves the best WER improvements at the 15-best depth with 2.9% absolute WER reduction (9.6% relative WER reduction). This is not very surprising since nearly 80% of the total WER improvement listed by the Oracle is hidden within the 20-best hypotheses. This strengthens our proposition about gaining substantial improvements by applying a strong post-processing mechanism at a small n-best depth. We believe that the 2.9% improvement is achieved due to the knowledge sources complementing each other in tandem and hence, a breakdown of the improvements from the various knowledge sources will not reflect their actual individual contributions but higher weights assigned to the phonetic and semantic features make them the main WER improvement contributors. We would also like to point out that the baseline WER is around 5-7% higher because we did not spend time tuning the pronounciation dictionary for the SWB task, nor did we include rover or any similar technique into the baseline and, the system was not tuned for accuracy while compromising on speed.

The use of the non-domain specific knowledge sources not only improves the WER for CSR as shown for the Switchboard task but also for goal based telephone speech (one of the reasons not to include special CSR features like incomplete sentences and disfluencies). Reference [9] presents a efficient procedure to create and tune grammars for telephone based directed dialog speech applications using only spoken user utterances. A necessary first step in this automation process is transcribing the user responses more accurately than has previously been possible. Using the same reranking setting as for the Switchboard task, we achieve 6.9% absolute WER reduction (14.47% relative reduction) by reranking a 30-best list for a set of 8013 user utterances for 4 prompts from 3 different applications.

4. CONCLUSIONS AND FUTURE WORK

This paper presented a n-best reranking mechanism that uses non-domain specific and more extensive additional phonetic, lexical, syntactic and semantic knowledge to improve WER in telephonic utterance recognition. Our future work includes increasing the coverage of the semantic model by including other content word predicates along with verbs into a more robust statistical argument relations identification and labeling framework. We also intend to devise a better and robust knowledge combination technique or possibly embed some of these knowledge sources directly into the LVCSR.

5. REFERENCES

[1] B. Pellom, *SONIC: The University of Colorado Continuous Speech Recognizer*, University of Colorado, May 2005.

[2] H. Erdogan, R. Sarikaya, Y. Gao, and M. Picheny, "Semantic structured language models," in *ICSLP*, 2002.

[3] S. Harabagiu, D. Moldovan, and J. Picone, "Open-domain voice-activated question answering," in *COLING*, 2002.

[4] M. Jeong, B. Kim, and G. G. Lee, "Using higher-level linguistic knowledge for speech recognition error correction in a spoken q/a dialog," in *HLT-NAACL Workshop on Higher Level Linguistic Information for Speech Processing*, 2004.

[5] S. Khudanpur and J. Wu, "Maximum entropy techniques for exploiting syntactic, semantic and collocational dependencies in language modeling," *CSL*, 2000.

[6] C. Chelba and F. Jelinek, "Exploiting syntactic structure for language modeling," in *COLING-ACL*, 1998.

[7] W. Wang, A. Stolcke, and M. P. Harper, "The use of a linguistically motivated language model in conversational speech recognition," in *ICASSP*, 2004.

[8] K. Hall and M. Johnson, "Attention shifting for parsing speech," in *ACL*, 2004.

[9] M. Balakrishna, D. Moldovan, and E. K. Cave, "Higher level phonetic and linguistic knowledge to improve asr accuracy and its relevance in interactive voice response systems," in *AAAI Workshop on SLU*, 2005.

[10] A. Ganapathiraju, *Support Vector Machines for Speech Recognition*, Ph.D. thesis, Mississippi State University, 2002.

[11] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus among words: Lattice-based word error minimization," in *Eurospeech*, 1999.

[12] E. Charniak, "Immediate-head parsing for language models," in *ACL*, 2001.

[13] P. Kingsbury, M. Palmer, and M. Marcus, "Adding semantic annotation to the penn treebank," in *HLT*, 2002.

[14] S. Pradhan, W. Ward, K. Hacioglu, J. H. Martin, and D. Jurafsky, "Shallow semantic parsing using support vector machines," in *HLT-NAACL*, 2004.

[15] S. Harabagiu, G. Miller, and D. Moldovan, "Wordnet 2 - a morphologically and semantically enhanced resource," in *SIGLEX*, 1999.