DIALOG-CONTEXT DEPENDENT LANGUAGE MODELING COMBINING N-GRAMS AND STOCHASTIC CONTEXT-FREE GRAMMARS

Kadri Hacioglu and Wayne Ward

Center for Spoken Language Research
University of Colorado at Boulder
E-mail: {hacioglu,whw}@cslr.colorado.edu

ABSTRACT

In this paper, we present our research on dialog dependent language modeling. In accordance with a speech (or sentence) production model in a discourse we split language modeling into two components; namely, dialog dependent concept modeling and syntactic modeling. The concept model is conditioned on the last question prompted by the dialog system and it is structured using n-grams. The syntactic model , which consists of a collection of stochastic context-free grammars one for each concept, describes word sequences that may be used to express the concepts. The resulting LM is evaluated by rescoring N-best lists. We report significant perplexity improvement with moderate word error rate drop within the context of CU Communicator System; a dialog system for making travel plans by accessing information about flights, hotels and car rentals.

1. INTRODUCTION

Statistical modeling of spoken language structure is crucial for the speech recognition and speech understanding components of dialog systems. Two broad statistical language models (LMs) that have been extensively studied are n-grams [1] and stochastic context free grammars (SCFGs) [2].

The standard n-gram LM tries to capture the structure of a spoken language by assigning probabilities to words conditioned on $n-1$ preceding words. The value of n is usually kept low (2 or 3) since (a) the number of parameters increases exponentially with n and (b) the training data is sparse, particularly, in early phases of sytem development. Therefore, standard n-gram LMs do not model longer distance correlations. They also do not take advantage of linguistic knowledge or structure.

A SCFG consists of a number of non-terminals, terminals, production rules and rule probabilites. It defines a stochastic formal language. It is possible to define SCFGs at two levels; namely, sentence level and phrase level. Sentence level SCFGs provide complete syntactic analysis across a sentence considering all words. They are expected to work very well for grammatical sentences (those covered by the grammar) but completely fail in sentences with ungrammatical construction. So, their use for spoken language applications is very limited. On the other hand, phrase level SCFGs focus on the syntax of sentence fragments. They allow partial parsing of sentences and are more appropriate for spoken language modeling.

SCFGs have properties complementary to n-grams. They are combined in various ways to obtain LMs with better perplexity and speech recognition/understanding performance [3, 4, 5, 6]. A promising approach is the use of semantically motivated phrase level SCFGs to parse a sentence into a sequence of concept (or semantic) tokens which are modeled using n-grams.

In this paper, we consider the language modeling problem within the framework of concept decoding (an integrated approach to speech recognition and understanding) based on the speech production model in a typical dialog. This framework uses a dialog context dependent LM with two components that we describe in Section 2. The idea of using dialog contextual knowledge to improve speech recognition and speeh undersatnding is not new [7]. Dialog dependent LMs have been recently investigated in [8, 9, 10, 11]. The method presented here is an extension of the work in [4], which was developed from ideas presented in [12], to a dialog dependent language modeling. We use the resulting LM to rescore N-best lists from our dialog system known as CU Communicator [13]. The rescoring scheme is a crude approximation to the integrated approach. We report significant perplexity improvement along with moderate improvement in word error rate after the N-best list rescoring.

The paper is organized as follows. Section 2 presents the integrated approach as a motivation to the use of dialog dependent language modeling. In Section 3, we explain an N-best rescoring scheme as a first order approximation to the integrated approach. Section 4 explains syntactic and semantic models in detail. Experimental results are presented in Section 5. Concluding remarks are made in the last section.

2. INTEGRATED APPROACH

The speech production model that we base our approach on is depicted in Figure 1. It is a slightly modifed version of the model in [14]. The user is assumed to have a specific goal that does not change throughout the dialog. According to the goal and the dialog context the user first picks a set of concepts with respective values and then use phrase generators associated with concepts to generate the word sequence. The word sequence is next mapped into a sequence of phones and converted into a speech signal by the user's vocal apparatus which we finally observe as a sequence of acoustic feature vectors.

The work is supported by DARPA through SPAWAR under grant #N66001-00-2-8906.

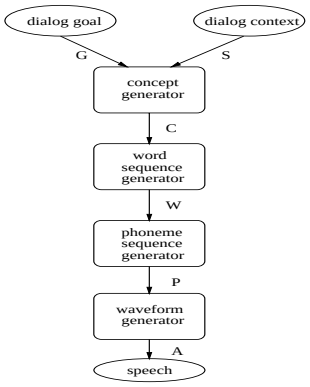

Fig. 1. A speech production model in a dialog

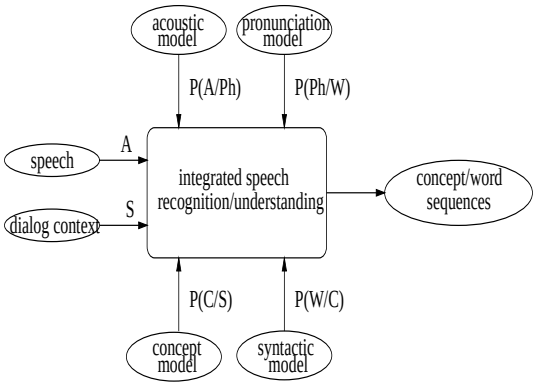

Fig. 2. Integrated speech recognition/understanding

The integrated approach, as depicted in Figure 2, is based on the speech production model, maximum a posteriori (MAP) optimization and Viterbi approximation:

$$C^*, W^* = \underset{C,W}{\text{argmax}} \ \max_{Ph} P(A/Ph)P(Ph/W)P(W/C)P(C/S) \quad (1)$$

where S is the dialog context, C is the sequence of concepts, W is the sequence of words, and Ph is the sequence of phones and A is the sequence of acoustic features. In (1) we identify four models:

- Concept model: $P(C/S)$
- Syntactic model : $P(W/C)$
- Pronunciation model: $P(Ph/W)$
- Acoustic model: $P(A/Ph)$

The concept model is the a priori probabilities of concept sequences conditioned on the dialog context. The syntactic model is the probability of word strings used to express a given concept. The pronunciation model gives the probabilities of possible phonetic realizations of a word. The acoustic model is the probability for the occurance of acoustic feature observations given phones.

Direct optimization of (1) is computationaly very demanding. For real time performance one ought to implement (1) in multiple stages at the expense of optimality. Since our major concern is the design, understanding and use of the concept and syntactic models with real-time performance we have chosen the simplest possible set-up to evaluate the models.

3. N-BEST LIST RESCORING

A quick and easy way of checking a new language model is rescoring of N-best lists from a speech recognizer that works with a relatively simple language model. The N-best list is a collection of sentences ranked according to their total acoustic and LM scores. Usually, rescoring is done by replacing the old LM scores with the new LM scores. Within the framework of Section 2, the N-best rescoring can be stated as the following MAP optimization:

$$C^*, W^* = \underset{C,W \in L_N}{\text{argmax}} \ p_A P(W/C)P(C/S) \quad (2)$$

where $p_A = P(A/Ph)P(Ph/W)$ is the acoustic probability from the first pass, L_N denotes the N-best list. If there is another mechanism to yield a unique concept sequence associated with W, say C_W, the N-best list rescoring in (2) will be:

$$W^* = \underset{W \in L_N}{\text{argmax}} \ p_A P(W/C_W)P(C_W/S) \quad (3)$$

Our work is based on (3). The rescoring scheme is illustrated in Figure 3. The parser provides C_W and $P(W/C_W)$ using SCFGs. The pool of LMs is used to calculate $P(C_W/S)$ and the best word sequence is output after rescoring.

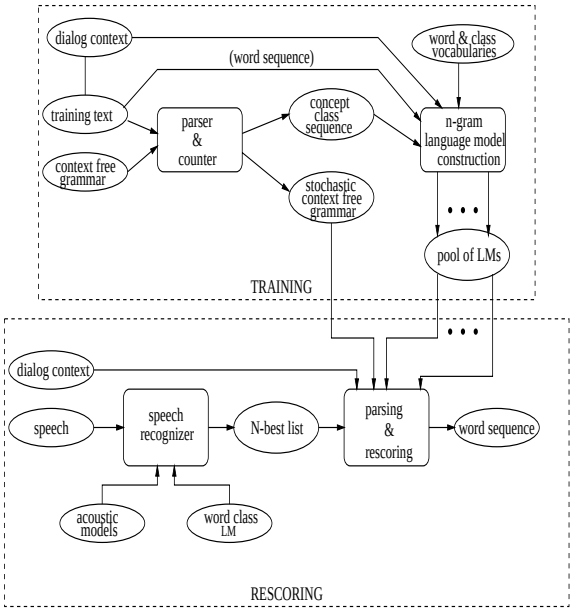

Fig. 3. Generic diagram of the system

<s> I WANT TO FLY FROM MIAMI FLORIDA TO SYDNEY AUSTRALIA ON OCTOBER FIFTH </s>
<s> [i_want] [depart_loc] [arrive_loc] [date] </s>

<s> I DON'T TO FLY FROM MIAMI FLORIDA TO SYDNEY AFTER AREA ON OCTOBER FIFTH </s>
<s> [Pronoun] [Contraction] [depart_loc] [arrive_loc] [after] [Noun] [date] </s>

Fig. 4. Examples of parsing into concepts and filler classes

4. CONCEPT AND SYNTACTIC MODELS

The concept models are conditioned on the dialog context. Although there are several ways to define a dialog context, we select the last question prompted as the dialog context. It is simple and yet strongly predictive and constraining.

The concepts are classes of phrases with the same meaning. Put differently, a concept class is a set of all phrases that may be used to express that concept (e.g. [i_want], [arrive_loc]). Those classes are augmented with single word, multiple word and a small number of broad (and unambigious) part of speech (POS) classes. In cases where the parser fails, we break the phrase into a sequence of words and tag them using this set of "filler" classes. Two examples in Figure 4 clearly illustrate the scheme.

The structure of the concept sequences is captured by an n-gram LM. We use two methods to condition the n-gram probabilities to the dialog context. In the first method we replace the sentence begin symbol <s> with the <dialog_context_name> and train a single n-gram LM with a lexicon that includes dialog context names as context cues in place of <s>. In the second method we train a seperate language model for each dialog context. Given the context S and $C = c_0 c_1 \cdots c_K, c_{K+1}$, the concept sequence probabilities are calculated as (for $n = 3$)

$$P(C/S) = \underset{K+1}{P}(c_1/ < S >)P(c_2/ < S >, c_1) \prod_{k=3} P(c_k/c_{k-2}, c_{k-1}) \quad \text{(method 1)}$$

$$P(C/S) = \underset{K+1}{P}(c_1/ < s >, S)P(c_2/ < s >, c_1, S) \prod_{k=3} P(c_k/c_{k-2}, c_{k-1}, S) \quad \text{(method 2)}$$

where c_0 and c_{K+1} are for the sentence-begin and sentence-end symbols, respectively.

Each class is written as a CFG and compiled into a stochastic recursive transition network (SRTN). The production rules are complete paths beginning from the start-node through the end-node in these nets. The probability of a complete path traversed through one or more SRTNs initiated by the top-level SRTN associated with the concept is the probability of the phrase that belongs to that concept. This probability is calculated as the multiplication of all arc probabilities that defines the path. That is,

$$\begin{aligned} P(W/C) &= \prod_{i=1}^{K} P(s_i/c_i) \\ &= \prod_{i=1}^{K} \prod_{j=1}^{M_i} P(r_j/c_i) \end{aligned}$$

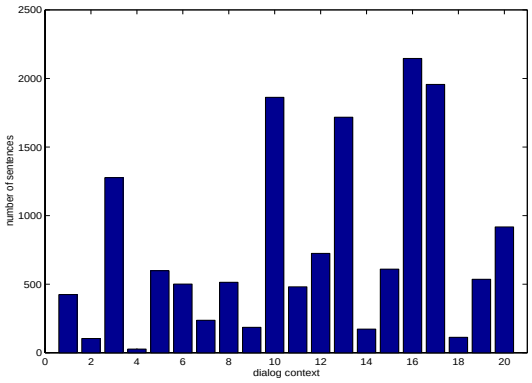

Fig. 5. Distrubition of sentences wrt dialog context

where s_i is a substring in $W = w_1, w_2..w_L = s_1, ..s_2, s_K$ ($K \leq L$) and $r_1, r_2, ...r_{M_i}$ are the production rules that construct s_i. The concept and rule sequences are assumed to be unique in the above equations. The parser uses heuristics to comply with this assumption.

The training procedure for the models are shown in Figure 3. SCFG and n-gram probabilities (including word based n-grams) are learned from a text corpus by simple counting. We believe that the degree of ambiguity in our models does not favor computationaly intensive stochastic training and parsing methods, thanks to our efforts to keep it as low as possible and the semantic-driven grammar.

5. EXPERIMENTAL RESULTS

The models were developed and tested in the context of the CU Communicator dialog system which is used for flight, hotel and rental car reservations [13]. The text corpus was divided into three parts as training, development and test sets with 15220, 450 and 770 sentences, respectively. The development set was used to optimize language weights and smoothing parameters. All results were reported using the test set. The average sentence length of the corpus was 4 words (end-of-sentence was treated as a word). We identified 20 dialog contexts and labeled each sentence with the associated dialog context. The distrubition of sentences across dialog contexts are shown in Figure 5.

We trained dialog dependent (DD) (for both method-1 and method-2) and dialog independent (DI) word, class and grammar based LMs. In all LMs n is set to 3. It must be noted that the DI class-based LM has served as the LM of the baseline system with 921 unigrams including 19 classes. The total number of the distinct words in the lexicon was 1681. The grammar-based LM had 199 concept and filler classes that completely cover the lexicon. The perplexity results are presented in Table 1.

Although the target LM is DD-2 SCFG 3-gram we present perplexities of all LMs from the least sophisticated (DI word 3-gram LM) to the most sophisticated (DD-2 SCFG 3-gram LM) to provide a guide for the choice of an LM. The perplexity improvement with the target LM compared to the simplest LM is 56% and

Table 1. Perplexity results

	DI	DD-1	DD-2
word 3-gram	25.1	19.8	23.1
class 3-gram	18.1	13.6	14.9
SCFG 3-gram	15.1	11.1	11.1

compared to the baseline LM (DI class 3-gram) is 39%. Although DD LMs with method 2 were expected to yield the best results, Table 1 clearly shows that DD LMs with method-1 gave the best results. The reasons are (a) fairly small average sentence length of the task and (b) data sparseness particularly in some dialog contexts (see Figure 5). Therefore, the DD language modeling using method-1 is of special interest for tasks where user utterances are quite short and a small corpus of sentences is available for training. We believe that the results will be reversed as more data comes in and we generalize the dialog contexts into a relatively smaller set. Note that the grammar based LM is the least sensitive to data sparseness making it useful particularly in early phases of system deployment.

We did some experiments using N-best lists from the baseline recognizer. We first determined the best possible performance in WER offered by N-best lists. This is done by picking the hypothesis with the lowest WER from each list. We report the results for several values of N in Table 2. They upperbound the performance gain possible from N-best list rescoring. We also present the rescoring results in Table 2. The language model used for rescoring is DD-2 SCFG LM. The result with N=10 (the best so far) amounts to 6.3% relative improvement in WER. This improvement is 26% of the improvement offered by the 10-best list.

Table 2. The best possible absolute WER drop in N-best list and the WER drop after N-best list rescoring: the baseline WER is 25.4%

N	1	10	20	50	100
Best	0.0	6.2%	7.3%	8.2%	8.7%
DD-2 SCFG	0.0	1.6%	1.4%	1.5%	1.5%

6. CONCLUSIONS

We have presented our recent work on language modeling using concept and syntactic models. The preliminary results show 39% relative improvement in perplexity and 6.3% relative improvement in WER after N-best rescoring with the target LM compared to the present class-based LM used in our Communicator system. Our work is in progress in several directions: (a) rescoring of word lattice, (b) clustering of dialog contexts, (c) interpolation of LMs and (d) assesment of the impact on concept accuracy.

7. REFERENCES

[1] L. Bahl, F. Jelinek, and Robert Mercer, "A maximum likelihood approach to continuous recognition," *IEEE Transactions on Pattern Analysis and Machine Inteligence*, vol. 5, no. 2, pp. 179–190, March 1983.

[2] J. K. Baker, "Trainable grammars for speech recognition," in *Speech Communications for th 97th Meeting of the Acoustical Society of America*, June 1979, pp. 31–35.

[3] M. Meteer, "Statistical language modeling combining n-gram and context-free grammars," in *International Conference of Acoustics, Speech, and Signal Processing*, June 1993, pp. 37–40.

[4] J. Gillett and W. Ward, "A language model combining trigrams and stochastic context-free grammars," in *5-th International Conference on Spoken Language Processing*, Sydney, Australia, 1998, pp. 2319–2322.

[5] B. Souvignier, A. Keller, B. Rueber, H.Schramm, and F. Seide, "The thoughtful elephant: Strategies for spoken dialog systems," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 1, pp. 51–62, January 2000.

[6] Y.Y. Wang, M. Mahajan, and X.Huang, "A unified context-free grammar and n-gram model for spoken language processing," in *International Conference of Acoustics, Speech, and Signal Processing*, Istanbul, Turkey, 2000, pp. 1639–1642.

[7] S. Young, "The MINDS system: Using context and dialogue to enhance speech recognition," in *Proceedings of DARPA Conference*, 1989, pp. 131–136.

[8] H. Niemann W. Eckert, F. Gallwitz, "Combining stochastic and linguistic language models for recognition of spontaneous speech," in *International Conference of Acoustics, Speech, and Signal Processing*, Atlanta, USA, 1996, pp. 423–426.

[9] C. Popovic and P. Baggia, "Specialized language models using dialogue predictions," in *International Conference of Acoustics, Speech, and Signal Processing*, Munich, Germany, April 1997, pp. 423–426.

[10] F. Wessel and A. Baader, "Robust dialogue-state dependent language modeling using leaving-one-out," in *International Conference of Acoustics, Speech, and Signal Processing*, Phoenix, USA, March 1999, pp. 741–744.

[11] F. Wessel and A. Baader, "A comparison of dialogue-state dependent language models," in *ESCA Workshop*, Kloster Irsee, Germany, June 1999, pp. 93–96.

[12] W. Ward and S. Young, "Flexible use of semantic constraints in speech recognition," in *International Conference of Acoustics, Speech, and Signal Processing*, June 1993, pp. 49–50.

[13] W. Ward and B. Pellom, "The CU communicator system," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, Keystone, Colorado, 1999.

[14] A. Keller, B. Rueber, F. Seide, and B.H. Tran, "PADIS - an automatic telephone switchboard and directory information system," *Speech Communication*, vol. 23, pp. 95–111, 1997.