Mixture Probabilistic Context-Free Grammar: An Improvement of a Probabilistic Context-Free Grammar Using Cluster-Based Language Modeling

Kenji Kita[†]

This paper proposes an improved probabilistic CFG (Context-Free Grammar), called the *mixture probabilistic CFG*, based on an idea of cluster-based language modeling. This model assumes that the language model parameters have different probability distributions in different topics or domains. In order to perform topic- or domain-dependent language modeling, we first divide the training corpus into a number of subcorpora according to their topics or domains, and then estimate separate probability distributions from each subcorpus. Therefore, a mixture probabilistic CFG has several different probability distributions for CFG productions. The language model probability of a sentence is calculated as the mixture of these probability distributions. The mixture probabilistic CFG enables us to make a context- or topic-dependent language model, and thus accurate language modeling would be possible. The proposed model was evaluated by calculating test-set perplexity using the ADD (ATR Dialogue Database) corpus and a Japanese intra-phrase grammar. The mixture probabilistic CFG had a test-set perplexity of 2.47/phone, while simple probabilistic CFG had a test-set perplexity of 2.77/phone. We also conducted speech recognition experiments using three language models, including pure CFG (without probabilities), simple probabilistic CFG, and the mixture probabilistic CFG. In our experiments, the mixture probabilistic CFG attained the best performance. The proposed model was also evaluated using sentence-level clustering. This evaluation used the dialogue corpus in which each utterance is annotated with an utterance type called IFT (Illocutionary Force Type). Using these IFTs, we divided the corpus into 9 clusters, and then estimated production probabilities from these clusters. Without IFT clustering, the perplexity was 2.18 per phone, but using IFT clustering, the perplexity was reduced to 1.82 per phone.

KeyWords: *language model, probabilistic context-free grammar, mixture probabilistic context-free grammar, cluster-based language model*

1 Introduction

Recently, probabilistic language models have been shown effective in many natural language applications. One such application is automatic speech recognition. Speech inherently

[†] Tokushima University, Faculty of Engineering

contains ambiguities and uncertainties that cannot be resolved by pure acoustic information. During recognition, many acoustically similar hypotheses are built. To effectively rank these hypotheses, the speech recognizer is required to rely on linguistic likelihood as well as acoustic likelihood. A probabilistic language model provides the basis for calculating linguistic likelihood.

One well-known probabilistic language model is a probabilistic context-free grammar (CFG), that is a grammar whose production rules have attached to them a probability of being used. These production probabilities are usually estimated from a training corpus under a probabilistic independent assumption, that the choice of a production rule is independent of the context. But, this simple assumption often results in a poor estimate of probability. Recently, more powerful language models beyond simple probabilistic CFGs have attracted considerable attention (Chitrao, M. V. and Grishman, R. 1990; Magerman, D. M. and Marcus, M. P. 1991; Black, E., Jelinek, F., Lafferty, J., Magerman, D. M., Mercer, R. and Roukos, S. 1992; Chiang, T-H. Lin, Y-C. and Su, K-Y. 1995); some of them take context-sensitive probabilities into account.

This paper will describe an improved probabilistic CFG, called the *mixture probabilistic CFG*, based on an idea of cluster-based language modeling. The basic idea of this model involves clustering a training corpus into a number of subcorpora, and then training probabilistic CFGs from these subcorpora. At the clustering, the similar linguistic objects (e.g., belonging to the same context, topic or domain) are formed into one cluster. The resulting probabilistic CFGs become context- or topic-dependent, and thus accurate language modeling would be possible.

This paper is organized as follows. Section 2 gives an overview of a probabilistic CFG. Section 3 describes a mixture probabilistic CFG. Section 4 contains evaluation experiments, including language model evaluation experiments from the viewpoint of perplexity reduction and speech recognition experiments. Section 5 describes another evaluation experiments using the IFT-annotated dialogue corpus. Finally, Section 6 presents our conclusions.

2 Probabilistic CFG: An Overview

A probabilistic CFG (Fujisaki, T., Jelinek, F., Cocke, J., Black, E. and Nishino, T. 1991) extends a CFG so that each production rule is of the form $\langle A \rightarrow \alpha, p \rangle$, where p is the conditional probability of A being rewritten into α . The probabilities of all A -productions (rules having A on the LHS) should sum to 1.

In the probabilistic CFG, the probability of a derivation can be computed as the product

of the probabilities of the rules used. Suppose that

$$S \xRightarrow{r_1} \gamma_1 \xRightarrow{r_2} \gamma_2 \xRightarrow{r_3} \cdots \xRightarrow{r_n} \gamma_n = w \quad (1)$$

is a derivation of w from the start symbol S , then the probability of this derivation D is given by

$$P(D) = \prod_{i=1}^n P(r_i). \quad (2)$$

The probability of a sentence w is the sum of the probabilities of all possible derivations for w .

$$P(w) = \sum_D P(D) \quad (3)$$

The production probabilities are estimated from a training corpus as follows:

Definition of Symbols

$\{B_1, B_2, \dots, B_I\}$... A set of training sentences.

$\{D_1^i, D_2^i, \dots, D_{n_i}^i\}$... A set of derivations for the i -th sentence B_i . Here, n_i represents the number of derivations for B_i .

$N_j^i(r)$... This function counts the number of rule occurrences (of its arguments) in the derivation D_j^i .

Training of the Probabilistic CFG

The conditional probabilities of rules in the probabilistic CFG were estimated using the following procedure (Fujisaki, T., Jelinek, F., Cocke, J., Black, E. and Nishino, T. 1991).

- (1) Make an initial guess of $P(A \rightarrow \alpha)$ such that $\sum_{\alpha} P(A \rightarrow \alpha) = 1$ holds.
- (2) Parse the i -th sentence B_i and get all the derivations for B_i .
- (3) Re-estimate $P(A \rightarrow \alpha)$ by the following formula.

$$\overline{P(A \rightarrow \alpha)} = \frac{\sum_i C_A^i(\alpha)}{\sum_i \sum_{\beta} C_A^i(\beta)} \quad (4)$$

where

$$C_A^i(\alpha) = \sum_j \left(\frac{P(D_j^i)}{\sum_k P(D_k^i)} N_j^i(A \rightarrow \alpha) \right) \quad (5)$$

- (4) Replace $P(A \rightarrow \alpha)$ with $\overline{P(A \rightarrow \alpha)}$ and repeat from step 2.

3 Mixture Probabilistic CFG

3.1 Cluster-Based Language Modeling

There are two different approaches for cluster-based language modeling. The first approach addresses the data sparseness problem. In probabilistic language modeling, model parameters are usually estimated according to their frequencies in a training corpus. However, since the amount of available data is limited, many events are infrequent and do not occur in the corpus. To circumvent this problem, the training data is clumped into a number of clusters, which are then used to smooth probabilities of occurrence for infrequent events. A class-based n -gram model (Brown, P. F., Pietra, V. J. D., deSouza, P. V., Lai, J. C. and Mercer, R. L. 1992) is a typical example of this approach.

The second approach aims to increase the model precision. The basic assumption in this approach is that the language model parameters have different probability distributions in different topics or domains. The training corpus contains texts from various kinds of topics or domains. This approach first divides the training corpus into a number of subcorpora according to their topics or domains, and then performs topic- or domain-dependent language modeling. Works (Carter 1994; Iyer, R., Ostendorf, M. and Rohlicek, J. R. 1994) belongs to this category.

3.2 Mixture Probabilistic CFG

A mixture probabilistic CFG is based on the second approach. In a conventional manner, production probabilities are estimated using the whole training data. In a mixture probabilistic CFG, however, we divide the training corpus into N clusters, and estimate separate probability distribution for each cluster. Thus, as a result, we have N probability distributions for the CFG.

Now suppose that the training corpus T is divided into N clusters $T_1, T_2, \dots, T_N$. That is,

$$T = T_1 \cup T_2 \cup \dots \cup T_N \quad (6)$$

$$T_i \cap T_j = \phi \quad (\text{if } i \neq j) \quad (7)$$

Let $P_i(S)$ denote the probability of sentence S using the probability distribution obtained from cluster T_i . Then, the mixture probabilistic CFG calculates the probability of S as follows:

$$P(S) = \sum_{i=1}^N q_i P_i(S) \quad (8)$$

In Equation 8, q_i is the probability of sentence S arising from cluster T_i and calculated as follows:

$$q_i = \frac{|T_i|}{\sum_j |T_j|} \quad (9)$$

Here, $|T_i|$ indicates the number of sentences in cluster T_i .

4 Evaluation Experiments

4.1 Corpus and Grammar

In our evaluation experiments, we used the ADD (ATR Dialogue Database) corpus (Ehara, T., Ogura, K. and Morimoto, T. 1990), which was created by ATR Interpreting Telephony Research Laboratories in Japan. The ADD corpus is a large structured database of dialogues collected from simulated telephone or keyboard conversations which are spontaneously spoken or typed in Japanese or English.

Currently, the ADD corpus contains textual data from two tasks (text categories); one consists of simulated dialogues between a secretary and participants at international conferences (Conference Task), and the other of simulated dialogues between travel agents and customers (Travel Task). In our experiments, we used the keyboard dialogues from the Conference Task.

In the experiments, we also used a Japanese intra-phrase grammar for the Conference

<start>	→	<bunsetu>
<bunsetu>	→	<interj>
<bunsetu>	→	<conj>
<bunsetu>	→	<np>
<bunsetu>	→	<vaux>
<bunsetu>	→	<quote>
.....		
<np>	→	<n-suffix>
<np>	→	<n-suffix> <p-k-wa>
<np>	→	<n-hutu> <p-kaku-ga>
.....		
<interj>	→	m o sh i m o sh i
.....		

Fig. 1 Example of CFG productions.

Task. This grammar does not describe a sentence structure, but it describes constraints inside Japanese phrases. Figure 1 shows some productions in our grammar.

In Figure 1, the grammar symbols quoted by $\langle \rangle$ indicate nonterminal symbols. The start symbol, indicated by $\langle \text{start} \rangle$, is rewritten into phrase category names. For example, $\langle \text{interj} \rangle$, $\langle \text{conj} \rangle$ and $\langle \text{np} \rangle$ are nonterminal symbols for interjection words, conjunctive phrases and noun phrases, respectively. Our grammar was written for phone-based speech recognition, thus terminal symbols are phone names.

Table 1 shows the size of the grammar and the training/evaluation data.

Table 1 Size of the grammar and the training/evaluation data.

Number of productions	2,590 rules
Number of words	1,591 words
Number of training data	34,301 phrases
Number of evaluation data	693 phrases

4.2 Corpus Clustering

Corpus clustering is required to derive probability distributions in a mixture probabilistic CFG. In our evaluation experiments, the clustering was conducted using phrase category names such as $\langle \text{interj} \rangle$, $\langle \text{conj} \rangle$ or $\langle \text{np} \rangle$. We first segmented the training corpus into phrases, and then assign a phrase category name to each phrase. Category assignment was carried out by analyzing each phrase using the intra-phrase grammar. In this way, the training corpus was divided into a number of clusters according to their phrase categories.

There is one thing that should be noted here. Since the parameters for the mixture probabilistic CFG are derived by statistical estimation from each cluster, the size of each cluster (the number of phrases belonging to each cluster) is largely responsible for the quality of the model. In other words, in order to estimate the reliable probabilities, each cluster must have enough data. In our experiments, the intra-phrase grammar had 109 phrase categories. However, after clustering based on these 109 categories, some clusters had very few data. For the reliable statistical estimation, clusters having fewer than 10 phrases (32 clusters in total) were merged into one cluster. As a result, we had 78 clusters obtained.

4.3 Evaluation Results

To evaluate the quality of a mixture probabilistic CFG, we calculated the *test-set perplexity* (Lee 1989). As a comparison, we also calculated the test-set perplexity of a simple probabilistic CFG. The test-set perplexity is the information-theoretic average branching of words along the test sentences (test-set), and is used as a measure of the difficulty of a recognition task relative to a given language model. In general, speech recognition performance is expected to increase as the test-set perplexity decreases. Thus, a language model with low perplexity is better.

As stated earlier, terminal symbols of the CFG are phone names. Therefore, we actually calculated the test-set perplexity per phone. A formula for the test-set perplexity per phone, PP , is given by:

$$PP = 2^{LP} \quad (10)$$

$$LP = -\frac{1}{N_W} \sum_{i=1}^{N_S} \log_2 P(S_i) \quad (11)$$

where N_S is the total number of phrases in the test-set, N_W is the total number of phones in all phrases, and $P(S_i)$ is the language model probability for the i -th phrase S_i . The results of perplexity measurements are summarized in Table 2. The mixture probabilistic CFG had a test-set perplexity of 2.47/phone, while simple probabilistic CFG had a test-set perplexity of 2.77/phone. This result supports the effectiveness of the mixture probabilistic CFG.

Table 2 Test-set perplexity

Simple probabilistic CFG	2.77 / phone
Mixture probabilistic CFG	2.47 / phone

4.4 Speech Recognition Experiments

We also conducted speech recognition experiments using three language models:

- (1) Pure CFG (without production probabilities),
- (2) Simple probabilistic CFG,
- (3) Mixture probabilistic CFG.

As the speech recognition system, we used the *HMM-LR system* (Kita, K., Kawabata, T. and Saito, H. 1989; Hanazawa, T., Kita, K., Nakamura, S., Kawabata, T. and Shikano, K. 1990), which is an integration of *hidden Markov models* (HMM) (Huang, X. D., Ariki, Y.

and Jack, M. A. 1990) and *generalized LR parsing* (Tomita 1991). The HMM-LR system is a syntax-directed continuous speech recognition system. The system outputs sentences that the grammar can accept.

The speech recognition experiments were conducted under the speaker-dependent condition, using discrete-type, context-independent HMMs without duration control. The results reported in Table 3 compare three language models in terms of phrase recognition performance. The mixture model attains the best performance.

Table 3 Phrase recognition performance

Pure CFG (without production probabilities)	83.6%
Simple probabilistic CFG	86.4%
Mixture probabilistic CFG	89.0%

5 Sentence-Level Evaluation Experiments

In the previous section, we have described evaluation experiments based on phrase-level clustering. This section describes another evaluation experiments, in which sentence-level clustering are performed based on discourse-level information.

In this evaluation, we used a dialogue corpus in which each utterance is annotated with an utterance type (Nagata, M. and Morimoto, T. 1994). The corpus used in the evaluation is what we call “Model Dialogues”, which consists of 10 dialogues with 225 sentences. Each utterance is annotated with IFT (Illocutionary Force Type), which is an abstraction of the speaker’s intention in terms of the type of action the speaker intends by the utterance. The following 9 IFTs are used in this corpus:

- (1) **phatic** ... phatic expressions such as greetings (e.g. Hello, Good-bye)
- (2) **expressive** ... idioms that express the speaker’s feeling (e.g. Thank you, You’re welcome)
- (3) **response** ... idiomatic responses and short answers (e.g. Yes, I see, That’s right)
- (4) **promise** ... the speaker commits himself to perform an action (e.g. I will send you a registration form)
- (5) **request** ... the speaker asks the hearer to perform an action (e.g. Please go to Kitaooji station by subway)
- (6) **inform** ... informative expressions (e.g. We are not giving any discount this time)
- (7) **questionif** ... Yes-No questions (e.g. Do you have the announcement of the confer-

ence?)

- (8) **questionref** ... WH questions (e.g. What should I do?)
- (9) **questionconf** ... confirmations (e.g. You have already transfered the registration fee, right?)

Our experiment used a sentential grammar for the conference registration task, which had 2,789 production rules including 1,591 words. As stated above, utterances in the dialogue corpus are annotated with their IFTs. Using these IFTs, we divided the dialogue corpus into 9 clusters, and then estimated production probabilities from these clusters. The resulting probabilistic CFG was evaluated by calculating perplexity. In this experiment, since the amount of available data is small (namely, 10 dialogues with 225 sentences), all data were used both in the training stage and evaluation stage. That is, perplexity measurement was performed against 225 sentences, which were also used in the probability estimation.

Without IFT clustering, the perplexity was 2.18 per phone, but using IFT clustering, the perplexity was reduced to 1.82 per phone. Although this evaluation was conducted using small-scale data, our result suggests the effectiveness of the mixture probabilistic CFG.

6 Conclusion

In this paper, we have proposed an improved probabilistic CFG, called the *mixture probabilistic CFG*, based on an idea of cluster-based language modeling. The basic idea of this model involves clustering a training corpus into a number of subcorpora, and then training probabilistic CFGs from these subcorpora. At the clustering, the similar linguistic objects (e.g., belonging to the same context, topic or domain) are formed into one cluster. The resulting probabilistic CFGs become context- or topic-dependent, and thus accurate language modeling would be possible. The proposed model was evaluated using the ADD (ATR Dialogue Database) corpus and a Japanese intra-phrase grammar. The effectiveness of the model was confirmed by perplexity reduction and speech recognition experiments.

Acknowledgement

The author would like to thank Masaaki Nagata (NTT Information and Communication Systems Laboratories) and Tsuyoshi Morimoto (ATR Interpreting Telecommunications Research Laboratories) for providing us with the IFT-annotated dialogue corpus.

Reference

Black, E., Jelinek, F., Lafferty, J., Magerman, D. M., Mercer, R. and Roukos, S. (1992). "To-

- wards history-based grammars: using richer models for probabilistic parsing.” *ACL-93*, pp. 31–37.
- Brown, P. F., Pietra, V. J. D., deSouza, P. V., Lai, J. C. and Mercer, R. L. (1992). “Class-based n -gram models of natural language.” *Computational Linguistics*, 18(4), 467–479.
- Carter, D. (1994). “Improving language models by clustering training sentences.” In *SRI Cambridge Technical Report, CRC-045*.
- Chiang, T.-H. Lin, Y.-C. and Su, K.-Y. (1995). “Robust learning, smoothing, and parameter tying on syntactic ambiguity resolution.” *Computational Linguistics*, 21(3), 321–349.
- Chitrao, M. V. and Grishman, R. (1990). “Statistical parsing of messages.” In *Proc. of DARPA Speech and Natural Language Workshop*, pp. 263–266.
- Ehara, T., Ogura, K. and Morimoto, T. (1990). “ATR dialogue database.” In *Proc. of the 1990 International Conference on Spoken Language Processing*, pp. 1093–1096.
- Fujisaki, T., Jelinek, F., Cocke, J., Black, E. and Nishino, T. (1991). “A probabilistic parsing method for sentence disambiguation.” In *Current Issues in Parsing Technology*. Tomita, M. (Ed.). 139–152. Kluwer Academic Publishers.
- Hanazawa, T., Kita, K., Nakamura, S., Kawabata, T. and Shikano, K. (1990). “ATR HMM-LR continuous speech recognition system.” In *Proc. of the 1990 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 53–56.
- Huang, X. D., Ariki, Y. and Jack, M. A. (1990). *Hidden Markov Models for Speech Recognition*. Edinburgh University Press.
- Iyer, R., Ostendorf, M. and Rohlicek, J. R. (1994). “Language modeling with sentence-level mixtures.” In *Proc. of ARPA Human Language Technology Workshop*, pp. 82–87.
- Kita, K., Kawabata, T. and Saito, H. (1989). “HMM continuous speech recognition using predictive LR parsing.” In *Proc. of the 1989 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 703–706.
- Lee, K.-F. (1989). *Automatic Speech Recognition: The Development of the SPHINX System*. Kluwer Academic Publishers.
- Magerman, D. M. and Marcus, M. P. (1991). “Parsing the Voyager domain using Pearl.” In *Proc. DARPA Speech and Natural Language Workshop*, pp. 231–236.
- Tomita, M. E. (1991). *Generalized LR Parsing*. Kluwer Academic Publishers.
- Nagata, M. and Morimoto, T. (1994). “An information-theoretic model of discourse for next utterance type prediction.” *Transactions of Information Processing Society of Japan*, 35(6), 1050–1061.

Kenji Kita: Kenji Kita received the B.S. degree in mathematics and the Ph.D degree in electrical engineering, both from Waseda University, Tokyo, Japan, in 1981 and 1992, respectively. From 1983 to 1987, he worked for the Oki Electric Industry Co. Ltd., Tokyo, Japan. From 1987 to 1992, he was a researcher at ATR Interpreting Telephony Research Laboratories, Kyoto, Japan. Since 1992, he has been with the Faculty of Engineering, Tokushima University, Tokushima, Japan, where he is currently an Associate Professor. He is a member of the Acoustical Society of Japan, the Information Processing Society of Japan, the Institute of Electronics, Information and Communication Engineers, and the Association for Computational Linguistics. His current research interests include speech recognition, natural language processing, and corpus linguistics.

(Received November 16, 1995)

(Revised April 1, 1996)

(Accepted June 14, 1996)