PROBABILISTIC LANGUAGE MODELING BASED ON MIXTURE PROBABILISTIC CONTEXT-FREE GRAMMAR

Kenji Kita and Tatsuya Iwasa

Faculty of Engineering
Tokushima University
Minami-josanjima, Tokushima 770, JAPAN

e-mail. {kita, iwasa}@is.tokushima-u.ac.jp

Abstract

This paper proposes an improved probabilistic CFG, called *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 effectiveness of the proposed model is confirmed both from perplexity reduction and speech recognition experiments.

1 Introduction

Recently, probabilistic language models have been shown effective in many natural language applications. One such application is automatic speech recognition. Speech inherently 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 [1, 2, 3, 4]; some of them take context-sensitive probabilities into account.

This paper will describe an improved probabilistic CFG, called *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. Finally, Section 5 presents our conclusions.

2 Probabilistic CFG: An Overview

A probabilistic CFG [5] extends a CFG so that each production rule is of the form $<A \rightarrow \alpha, p>$, 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

$$S \stackrel{r_1}{\Longrightarrow} \gamma_1 \stackrel{r_2}{\Longrightarrow} \gamma_2 \stackrel{r_3}{\Longrightarrow} \cdots \stackrel{r_n}{\Longrightarrow} \gamma_n = w \tag{1}$$

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

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

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

$$P(w) = \sum_{D} P(D) \tag{3}$$

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

Definition of Symbols

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

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

$N_j^i(r)$ $\cdots$ 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 [5].

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.

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

where

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

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 [6] 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 [7, 8] 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 clustsers $T_1, T_2, \cdots, T_N$. That is,

$$T = T_1 \cup T_2 \cup \cdots \cup T_N \tag{6}$$

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

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:

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

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

$$q_i = \frac{|T_i|}{\sum_j |T_j|} \tag{9}$$

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

<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
	.....	

Figure 1: Example of CFG productions.

In Figure 1, the grammar symbols quoted by <> indicate nonterminal symbols. The start symbol, indicated by <start>, is rewritten into phrase category names. For example, <inter>, <conj> and <np> are nonterminal symbols for interjection words, conjunctional phrases and noun phrases, respectively. Our grammar was written for phone-based speech recognition, thus terminal symbols were 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
Number of words	1,591
Number of training data	34,301
Number of evaluation data	693

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 <interj>, <conj> or <np>. 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 the intra-phrase grammar. In this way, the training corpus was devided 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 reponsible 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* [10]. 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 were 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:

$$PP = 2^{LP} \tag{10}$$

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

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, which 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:

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

As the speech recognition system, we used the *HMM-LR system* [11, 12], which is an integration of *hidden Markov models* (HMM) [13] and *generalized LR parsing* [14]. 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 Conclusion

This paper proposed an improved probabilistic CFG, called *mixture probabilistic CFG*, based on an idea of cluster-based language modeling. The effectiveness of the proposed model was confirmed by perplexity reduction and speech recognition experiments.

References

[1] Su, K. Y. and Chang, J. S.: "Semantic and Syntactic Aspects of Score Function", *Proc. COLING-88*, pp. 642-644 (1992).

[2] Chitrao, M. V. and Grishman, R.: "Statistical Parsing of Messages", *Proc. DARPA Speech and Natural Language Workshop*, pp. 263-266 (1990).

[3] Magerman, D. M. and Marcus, M. P.: "Parsing the Voyager Domain Using Pearl", *Proc. DARPA Speech and Natural Language Workshop*, pp. 231-236 (1991).

[4] Black, E., Jelinek, F., Lafferty, J., Magerman, D. M., Mercer, R. and Roukos, S.: "Towards History-based Grammars: Using Richer Models for Probabilistic Parsing", *Proc. DARPA Speech and Natural Language Workshop* (1992).

[5] Fujisaki, T., Jelinek, F., Cocke, J., Black, E. and Nishino, T.: "A Probabilistic Parsing Method for Sentence Disambiguation", In *Current Issues in Parsing Technology*, Tomita, M. (Ed.), pp. 139-152, Kluwer Academic Publishers (1991).

[6] Brown, P. F., Pietra, V. J. D., deSouza, P. V., Lai, J. C. and Mercer, R. L.: "Class-based n-gram Models of Natural Language", *Computational Linguistics*, Vol. 18, No. 4, pp. 467-479 (1992).

[7] Carter, D.: "Improving Language Models by Clustering Training Sentences", (1994).

[8] Iyer, R., Ostendorf, M. and Rohlicek, J. R.: "Language Modeling with Sentence-Level Mixtures", *Proc. of the Human Language Technology Workshop*, pp. 82-87 (1994).

[9] Ehara, T., Ogura, K. and Morimoto, T.: "ATR dialogue database", *Proc. of the 1990 International Conference on Spoken Language Processing*, pp. 1093-1096 (1990).

[10] Lee, K. F.: *Automatic Speech Recognition: The Development of the SPHINX System*, Kluwer Academic Publishers (1989).

[11] Kita, K., Kawabata, T. and Saito, H.: "HMM Continuous Speech Recognition Using Predictive LR Parsing", *Proc. ICASSP-89*, pp. 703-706 (1989).

[12] Hanazawa, T., Kita, K., Nakamura, S., Kawabata, T. and Shikano, K.: "ATR HMM-LR Continuous Speech Recognition System", *Proc. ICASSP-90*, pp. 53-56 (1990).

[13] Huang, X. D., Ariki, Y. and Jack, M. A.: *Hidden Markov Models for Speech Recognition*, Edinburgh University Press (1990).

[14] Tomita, M. (Ed.): *Generalized LR Parsing*, Kluwer Academic Publishers (1991).