A Preference-first Language Processor Integrating the Unification Grammar and Markov Language Model for Speech Recognition Applications

Lee-Feng Chien, K. J. Chen** and Lin-Shan Lee***

* Dept. of Computer Science and Information Engineering,
National Taiwan University,Taipei, Taiwan, Rep. of China, Tel: (02) 362-2444.
** The Institute of Information Science, Academia Sinica, Taipei, Taiwan, Rep. of China.

ABSTRACT

A language processor is to find out a most promising sentence hypothesis for a given word lattice obtained from acoustic signal recognition. In this paper a new language processor is proposed, in which unification grammar and Markov language model are integrated in a word lattice parsing algorithm based on an augmented chart, and the island-driven parsing concept is combined with various preference-first parsing strategies defined by different construction principles and decision rules. Test results show that significant improvements in both correct rate of recognition and computation speed can be achieved .

1. Introduction

In many speech recognition applications, a word lattice is a partially ordered set of possible word hypotheses obtained from an acoustic signal processor. The purpose of a language processor is then, for an input word lattice, to find the most promising word sequence or sentence hypothesis as the output (Hayes, 1986; Tomita, 1986; O'Shaughnessy, 1989). Conventionally either grammatical or statistical approaches were used in such language processors. However, the high degree of ambiguity and large number of noisy word hypotheses in the word lattices usually make the search space huge and correct identification of the output sentence hypothesis difficult, and the capabilities of a language processor based on either grammatical or statistical approaches alone were very often limited. Because the features of these two approaches are basically complementary, Derouault and Merialdo (Derouault, 1986) first proposed a unified model to combine them. But in this model these two approaches were applied primarily separately, selecting the output sentence hypothesis based on the product of two probabilities independently obtained from these two approaches.

In this paper a new language processor based on a recently proposed augmented chart parsing algorithm (Chien, 1990a) is presented, in which the grammatical approach of unification grammar (Sheiber, 1986) and the statistical approach of Markov language model (Jelinek, 1976) are properly integrated in a preference-first word lattice parsing algorithm. The augmented chart (Chien, 1990b) was extended from the conventional chart. It can represent a very complicated word lattice, so that the difficult word lattice parsing problem can be reduced to essentially a well-known chart parsing problem. Unification grammars, compared with other grammatical approaches, are more declarative and can better integrate syntactic and semantic information to eliminate illegal combinations; while Markov language models are in general both effective and simple. The new language processor proposed in this paper actually integrates the unification grammar and the Markov language model by a new preference-first parsing algorithm with various preference-first parsing strategies defined by different constituent construction principles and decision rules, such that the constituent selection and search directions in the parsing process can be more appropriately determined by Markovian probabilities, thus rejecting most noisy word hypotheses and significantly reducing the search space. Therefore the global structural synthesis capabilities of the unification grammar and the local relation estimation capabilities of the Markov language model are properly integrated. This makes the present language processor not sensitive at all to the increased number of noisy word hypotheses in a very large vocabulary environment. An experimental system for Mandarin speech recognition has been implemented (Lee, 1990) and tested, in which a very high correct rate of recognition (93.8%) was obtained at a very high processing speed (about 5 sec per sentence on an IBM PC/AT). This indicates significant improvements as compared to previously proposed models. The details of this new language processor will be presented in the following sections.

2. The Proposed Language Processor

The language processor proposed in this paper is shown in Fig. 1, where an acoustic signal preprocessor is included to form a complete speech recognition system. The language processor consists of a language model and a parser. The language model properly integrates the unification grammar and the Markov language model, while the parser is defined based on the augmented chart and the preference-first parsing algorithm. The input speech signal is first processed by the acoustic signal preprocessor; the corresponding word lattice will thus be generated and constructed onto the augmented chart. The parser will then proceed to build possible constituents from the word lattice on the augmented chart in accordance with the language model and the preference-first parsing algorithm. Below, except the preference-first parsing algorithm presented in detail in the next section, all of other elements are briefly summarized.

The Language Model

The goal of the language model is to participate in the selection of candidate constituents for a sentence to be identified. The proposed language model is composed of a PATR-II-like unification grammar (Sheiber, 1986; Chien, 1990a) and a first-order Markov language model (Jelinek, 1976) and thus, combines many features of the grammatical and statistical language modeling approaches. The PATR-II-like unification grammar is used primarily to distinguish between well-formed, acceptable word sequences against ill-formed ones, and then to represent the structural phrases and categories, or to find the intended meaning depending on different applications. The first-order Markov language model, on the other hand, is used to guide the parser toward correct search directions, such that many noisy word hypotheses can be rejected and many unnecessary constituents can be avoided, and the most promising sentence hypothesis can thus be easily found. In this way the weakness in either the PATR-II-like unification grammar (Sheiber, 1986), e.g., the heavy reliance on rigid linguistic information, or the first-order Markov language model (Jelinek, 1976), e.g., the need for a large training corpus and the local prediction scope can also be effectively remedied.

The Augmented Chart and the Word Lattice Parsing Scheme

Chart is an efficient and widely used working structure in many natural language processing systems (Kay, 1980; Thompson, 1984), but it is basically designed to parse a sequence of fixed and known words instead of an ambiguous word lattice. The concept of the augmented chart has recently been successfully developed such that it can be used to represent and parse a word lattice (Chien, 1990b). Any given input word lattice for parsing can be represented by the augmented chart through a mapping procedure, in which a minimum number of vertices are used to indicate the end points for all word hypotheses in the lattice, and an inactive edge is used to represent every word hypotheses. Also, specially designed jump edges are constructed to link some edges whose corresponding word hypotheses can possibly be connected but themselves are physically separated in the chart. In this way the basic operation of a chart parser can thus be properly performed on a word lattice. The difference is that two separated edges linked by a jump edge can also be combined as long as the required condition is satisfied. Note that in such a scheme, every constituents (edge) will be constructed only once, regardless of the fact that it may be shared by many different sentence hypotheses.

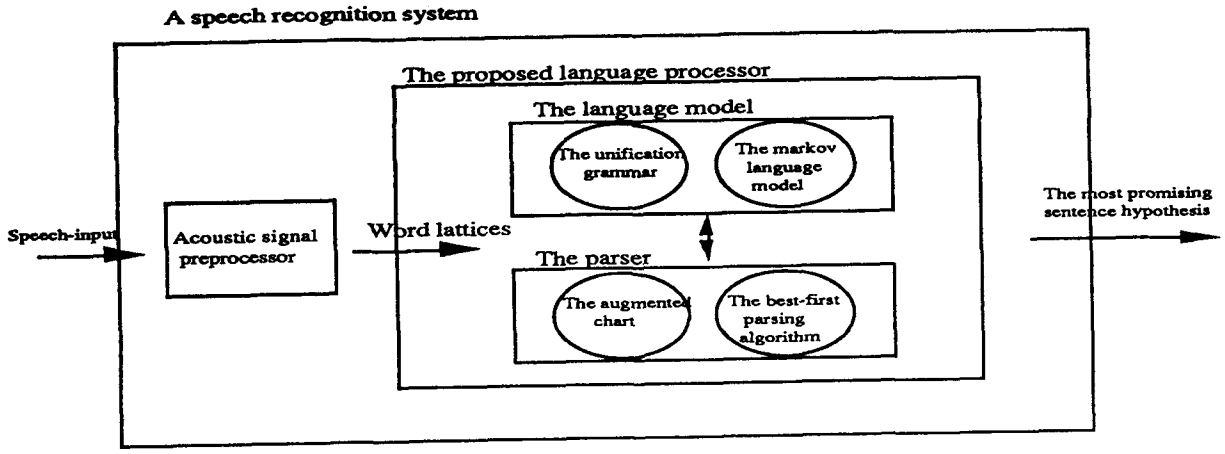

Fig. 1 An abstract diagram of the proposed language processor.

3. The Preference-first Parsing Algorithm

The preference-first parsing algorithm is developed based on the augmented chart summarized above, so that the difficult word lattice parsing problem is reduced to essentially a well-known chart parsing problem. This parsing algorithm is a general algorithm, in which various preference-first parsing strategies defined by different construction principles and decision rules can be combined with the island-driven parsing concept, so that the constituent selection and search directions can be appropriately determined by Markovian probabilities, thus rejecting many noisy word hypotheses and significantly reducing the search space. In this way, not only can the features of the grammatical and statistical approaches be combined, but the effects of the two different approaches are reflected and integrated in a single algorithm such that overall performance can be appropriately optimized. Below, more details about the algorithm will be given.

Probability Estimation for Constructed Constituents

In order to make the unification-based parsing algorithm also capable of handling the Markov language model, every constructed constituent has to be assigned a probability. In general, for each given constituent C a probability $P(C) = P(W_C)$ is assigned, where W_C is the component word hypothesis sequence of C and $P(W_C)$ can be evaluated from the Markov language model. Now, when an active constituent A and an inactive constituent I form a new constituent N, the probability P(N) can be evaluated from probabilities P(A) and P(I). Let W_n, W_a, W_i be the component word hypothesis sequences of N, A, and I respectively. Without loss of generality, assume A is to the left of I, thereby Wn = WaWi = $w_{a1},\ldots,w_{am},w_{i1},\ldots,w_{in}$, where w_{ak} is the k-th word hypothesis of Wa and w_{ik} the k-th word hypothesis of Wi. Then,

$$P(W_n) = P(WaWi) = P(w_{a1}) * \prod_{2 \le k \le m} P(w_{ak}|w_{ak-1}) * P(w_{i1}|w_{am}) * \prod_{2 \le k \le n} P(w_{ik}|w_{ik-1})$$
$$= P(W_a)*P(W_i)*\{P(w_{i1}|w_{am})/P(w_{i1})\}.$$

This can be easily evaluated in each parsing step.

The Preference-first Construction Principles and Decision Rules

Since P(C) is assigned to every constituent C in the augmented chart, various parsing strategies can be developed for the preference-first parsing algorithm for different applications. For example, there can be various construction principles to determine the order of constituent construction for all possible candidate constituents. There can also be various decision rules to choose the output sentence among all of the constructed sentence constituents. Some examples for such construction principles and decision rules are listed in the following.

Example Construction principles:

random principle: at any time randomly select a candidate constituent to be constructed

probability selection principle: at any time the candidate constituent with the highest probability will be constructed first

length selection principle: at any time the candidate constituent with the largest numt component word hypotheses will be constructed first

length/probability selection principle: at any time the candidate constituent with the highest probability among those with the largest number of component word hypotheses will be constructed first

Example Decision rules:

highest probability rule: after all grammatical sentence constituents have been found, one with the highest probability is taken as the result

first-1 rule: the first grammatical sentence constituent obtained during the course of parsing is taken as the result

first-k rule: the sentence constituent with the highest probability among the first k constructed grammatical sentence constituents obtained during the course of parsi is taken as the result

The performance of these various construction principles and decision rules will be discussed in Sections 5 and 6 based on experimental results.

4. The Experimental System

An experimental system based on the proposed language processor has been developed and tested on a small lexicon, a Markov language model, and a simple set of unification grammar rules for the Chinese language, although the present model is in fact language independent. The system is written in C language and performed on an IBM PC/AT.

The lexicon used has a total of 1550 words. They are extracted from the primary school Chinese text books currently used in Taiwan area, which are believed to cover the most frequently used words and most of the syntactic and semantic structures in the everyday Chinese sentences. Each word stored in the lexicon (word entry) contains such information as the word name, the pronunciations (the phonemes), the lexical categories and the corresponding feature structures. Information contained in each word entry is relatively simple except for the verb words, because verbs have complicated behavior and will play a central role in syntactic analysis. The unification grammar constructed includes about 60 rules. It is believed that these rules cover almost all of the sentences used in the primary school Chinese text books. The Markov language model is trained using the primary school Chinese text books as training corpus. Since there are no boundary markers between adjacent words in written Chinese sentences, each sentence in the corpus was first segmented into a corresponding word string before used in the model training. Moreover, the test data include 200 sentences randomly selected from 20 articles taken from several different magazines, newspapers and books published in Taiwan area. All the words used in the test sentences are included in the lexicon.

5. Test Results (I) -- Initial Preference-first Parsing Strategies

The present preference-first language processor is a general model on which different parsing strategies defined by different construction principles and decision rules can be implemented. In this and the next sections, several attractive parsing strategies are proposed, tested and discussed under the test conditions presented above. Two initial tests, test I and II, were first performed to be used as the baseline for comparison in the following. In test I, the conventional unification-based grammatical analysis alone is used, in which all the sentence hypotheses obtained from the word lattice were parsed exhaustively and a grammatical sentence constituent was selected randomly as the result; while in test II the first-order Markov modeling approach alone is used, and a sentence hypothesis with the highest probability was selected as the result regardless of the grammatical structure. The correct rate of recognition is defined as the averaged percentage of the correct words in the output sentences. The correct rate of recognition and the approximated average time required are found to be 73.8% and 25 sec for Test I , as well as 82.2% and 3 sec for Test II, as indicated in the first two rows of Table 1. In all the following parsing strategies, both the unification grammar and the Markov language model will be integrated in the language model to obtain better results.

The parsing strategy 1 uses the random selection principle and the highest probability rule (as listed in Section 3), and the entire word lattice will be parsed exhaustively. The total number of constituents constructed during the course of parsing for each test sentence are also recorded. The results show that the correct rate of recognition can be as high as 98.3%. This indicates that the language processor based on the integration of the unification grammar and the Markov language model can in fact be very reliable. That is, most of the interferences due to the noisy word hypotheses are actually rejected by such an integration. However, the computation load required for such an exhaustive parsing strategy turns out to be very high (similar to that in Test I), i.e., for each test sentence in average 305.9 constituents have to be constructed and it takes about 25 sec to process a sentence on the IBM PC/AT. Such computation requirements will make this strategy practically difficult for many applications. All these test data together with the results for the other three parsing strategies 2-4 are listed in Table 1 for comparison.

The basic concept of parsing strategy 2 (using the probability selection principle and the first-1 rule, as listed in Section 3) is to use the probabilities of the constituents to select the search direction such that significant reduction in computation requirements can be achieved. The test results (in the fourth row of Table 1) show that with this strategy for each test sentence in average only 152.4 constituents are constructed and it takes only about 12 sec to process a sentence on the PC/AT, and the high correct rate of recognition of parsing strategy 1 is almost preserved, i.e., 96.0%. Therefore this strategy represents a very good tradeoff, i.e., the computation requirements are reduced by a factor of 0.50 (the constituent reduction ratio in the last second column of Table 1 is the ration of the average number of built constituents to that of Strategy 1), while the correct rate is only degraded by 2.3%. However, such a speed (12 sac for a sentence) is still very low especially if real-time operation is considered.

6. Test Results (II) -- Improved Best-first Parsing Strategies

In a further analysis all of the constituents constructed by parsing strategy 1 were first divided into two classes: correct constituents and noisy constituents. A correct constituent is a constituent without any component noisy word hypothesis; while a noisy constituent is a constituent which is not correct. These two classes of constituents were then categorized according to their length (number of word hypotheses in the constituents). The average probability values for each category of correct and noisy constituents were then evaluated. The results are plotted in Fig. 2, where the vertical axis shows the average probability values and the horizontal axis denotes the length of the constituent. Some observations can be made as in the following. First, it can be seen that the two curves in Fig. 2 apparently diverge, especially for longer constituents, which implies that the Markovian probabilities can effectively discriminate the noisy constituents against the correct constituents (note that all of these constituents are grammatical), especially for longer constituents. This is exactly why parsing strategy 1 and 2 can provide very high correct rates. Furthermore, Fig. 2 also shows that in general the probabilities for shorter constituents would usually be much higher than those for longer constituents. This means with parsing strategy 2 almost all short constituents, no matter noisy or

correct, would be constructed first, and only those long noisy constituents with lower probability values can be rejected by the parsing strategy 2. This thus leads to the parsing strategies 3 and 4 discussed below.

In parsing strategy 3 (using the length/probability selection principle and First-1 rule, as listed in Section 3), the length of a constituent is considered first, because it is found that the correct constituents have much better chance to be obtained very quickly by means of the Markovian probabilities for longer constituents than shorter correct constituents, as discussed in the above. In this way, the construction of the desired constituents would be much more faster and very significant reduction in computation requirements can be achieved. The test results in the fifth row of Table 1 show that with this strategy in average only 70.2 constituents were constructed for a sentence, a constituent reduction ratio of 0.27 is found, and it takes only about 4 sec to process a sentence on PC/AT, which is now very close to real-time. However, the correct rate of recognition is seriously degraded to as low as 85.8%, apparently because some correct constituents have been missed due to the high speed construction principle. Fortunately, after a series of experiments, it was found that in this case the correct sentences very often appeared as the second or the third constructed sentences, if not the first. Therefore, the parsing strategy 4 is proposed below, in which everything is the same as parsing strategy 3 except that the first-1 decision rule is replaced by the first-3 decision rule. In other words, those missed correct constituents can very possibly be picked up in the next few steps, if the final decision can be slightly delayed.

The test results for parsing strategy 4 listed in the sixth row of Table 1 show that with this strategy the correct rate of recognition has been improved to 93.8% and the computation complexity is still close to that of parsing strategy 3, i.e., the average number of constructed constituents for a sentence is 91.0, it takes about 5 sec to process a sentence, and a constituent reduction ratio of 0.29 is achieved. This is apparently a very attractive approach considering both the accuracy and the computation complexity. In fact, with the parsing strategy 4, only those noisy word hypotheses which both have relatively high probabilities and can be unified with their neighboring word hypotheses can cause interferences. This is why the noisy word hypothesis interferences can be reduced, and the present approach is therefore not sensitive at all to the increased number of noisy word hypotheses in a very large vocabulary environment. Note that although intuitively the integration of grammatical and statistical approaches would imply more computation requirements, but here in fact the preference-first algorithm provides correct directions of search such that many noisy constituents are simply rejected and the reduction of the computation complexity makes such an integration also very attractive in terms of computation requirements.

7. Concluding Remarks

In this paper, we have proposed an efficient language processor for speech recognition applications, in which the unification grammar and the Markov language model are properly integrated in

	construction principles	decision rules	Correct rates of recognition	Number of built constituents	Constituent reduction ratios	Approximated average time required (Sec/Sentence)
Test I (Unification grammar only)	———	———	73.8 %	305.9	1.00	25
Test II (Markov language model only)	———	———	82.2 %	———	———	3
parsing strategy 1	the random selection principle	the highest probability	98.3 %	305.9	1.00	25
parsing strategy 2	the probability selection principle	First-1 rule	96.0 %	152.4	0.50	12
parsing strategy 3	the length/probability selection principle	First-1 rule	85.8 %	70.2	0.27	4
parsing strategy 4	the length/probability selection principle	First-3 rule	93.8 %	91.0	0.29	5

Table 1 Test results for the two initial tests and four parsing strategies.

a preference-first parsing algorithm defined on an augmented chart. Because the unification-based analysis eliminates all illegal combinations and the Markovian probabilities of constituents indicates the correct direction of processing, a very high correct rate of recognition can be obtained. Meanwhile, many unnecessary computations can be effectively eliminated and very high processing speed obtained due to the significant reduction of the huge search space. This preference-first language processor is quite general, in which many different parsing strategies defined by appropriately chosen construction principles and decision rules can be easily implemented for different speech recognition applications.

References:

Chien, L. F., Chen, K. J. and Lee, L. S. (1990b). An Augmented Chart Parsing Algorithm Integrating Unification Grammar and Markov Language Model for Continuous Speech Recognition. *Proceedings of the IEEE 990 International Conference on Acoustics, Speech and Signal Processing*, Albuquerque, NM, USA, Apr. 1990.

Chien, L. F., Chen, K. J. and Lee, L. S. (1990a). An Augmented Chart Data Structure with Efficient Word Lattice Parsing Scheme in Speech Recognition Applications. To appear on *Speech Communication.*, also in *Proceedings of the 13th International Conference on Computational Linguistics*, July 1990, pp. 60-65.

Derouault A. and Merialdo B. (1986). Natural Language Modeling for Phoneme-to-Text Transcription, *IEEE Trans. on PAMI*, Vol. PAMI-8, pp. 742-749.

Hayes, P. J. et al. (1986). Parsing Spoken Language:A Semantic Caseframe Approach. *Proceedings of the 11th International Conference on Computational Linguistics*, University of Bonn, pp. 587-592.

Jelinek, F. (1976). Continuous Speech Recognition by Statistical Methods, *Proc. IEEE*, Vol. 64(4), pp. 532-556, Apr. 1976.

Kay M. (1980). Algorithm Schemata and Data Structures in Syntactic Processing. Xerox Report CSL-80-12, pp. 35-70, Pala Alto.

Lee, L. S. et al. (1990). A Mandarin Dictation Machine Based Upon A Hierarchical Recognition Approach and Chinese Natural Language Analysis, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 12, No. 7. July 1990, pp. 695-704.

O'Shaughnessy, D. (1989). Using Syntactic Information to Improve Large Vocabulary Word Recognition, *ICASSP'89*, pp. 715-718.

Sheiber, S. M. (1986). An Introduction to Unification-Based Approaches to Grammar. *University of Chicago Press*, Chicago.

Thompson, H. and Ritchie, G. (1984). Implementing Natural Language Parsers, in *Artificial Intelligence, Tools, Techniques, and Applications*, O'shea, T. and Elsenstadt, M. (eds), Harper&Row, Publishers, Inc.

Tomita, M. (1986). An Efficient Word Lattice Parsing Algorithm for Continuous Speech Recognition. *Proceedings of the 1986 International Conference on Acoustic, Speech and Signal Processing*, pp. 1569-1572.

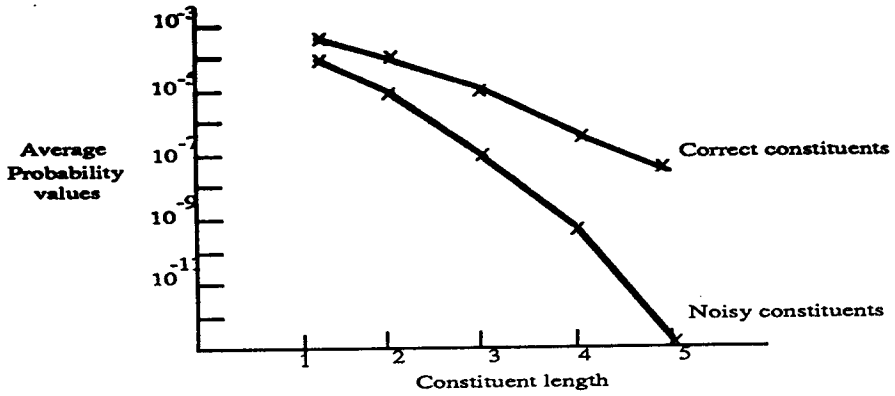

Fig. 2 The average probability values for the correct and noisy constituents with different lengths constructed by parsing strategy 1.