COMPUTER-AIDED GRAMMAR ACQUISITION IN THE CHINESE UNDERSTANDING SYSYTEM CUSAGA•

Lu Ruqian Liu Yinghui Li Xiaobin Institute of Mathematics, Acadcmia Sinica, Beijing

ABSTRACT

CALAS is a subsystem for acquiring semantic grammars to be used in CUSAGA which can understand technical Chinese texts and extract knowledge from them. The semantic grammar is acquired in a semi—automatic way under the guidance of the user. CUSAGA is implemented on UV68000 with about 12000 PASCAL lines. This paper gives a short overview on the architcclure and functions of CUSAGA and a more detailed

2 T<u>HE OVERAL</u>L STRU<u>CTURE OF CUS</u>AGA

The major components of CUSAGA arc two subsystems: IUC and CALAS. IUC is composed of the modules for segmenting, parsing, semantic analysis and understanding. With some initial knowledge stored in the knowledge base of CUSAGA, IUC is capable of accepting Chinese technical materials, parsing them and doing semantic analysis to create internal representations. The subsystem CALAS is composed of the modules for segmenting, parsing. learning morphology and learning gra mmar. Guided by the user, CALAS can acquire morphological and grammar rules from a set of legal Chinese sentences and put them into the knowledge base of CUSAGA, which will then be used by IUC. The two subsystems share some common submodules and arc linked up by the knowledge base of CUSAGA. Since IUC will be explained in more detail in another future paper, it is only briefly described here and the major part of the present paper is dedicated to CALAS.

discussion about the working principles and techniques adopted by CALAS is presented.

1 INTRODUCTION

Chinese is different from Western languages in pronunciation, written form, vocabulary and grammar. In Chinese, the logical relationship between character and character, or word and word, or sentence and sentence, depends on their order, their meaning and some function words. Chinese words have no inflection, no matter in what positions they appear in the sentence. A written Chinese sentence is a line of consecutive Chinese characters which contains no marks helping to segment words and phrases. All these peculiarities have made Chinese very difficult for computer processing.

Research works on computer-aided Chinese understanding were very rare until the beginning of seventies. During last 10 years, more and more Chinese computer scientists became interested in this topic. One of the most important goals has been the development of Chinese understanding interfaces, see [Yao, 1985], [Zhao, 1987], [Xie, 1988]. At the same time, several theoretical models for understanding Chinese were proposed, e.g., [Gao, 1985], [Wang, 1983], [Wang, 1987]. One shortcoming we found among these research works is that complicated models arc usually not implemented, and most of the implemented systems use relative simple models. Another problem is that almost all of the implemented systems are qucry interfaces, or small systems only for demonstration purpose. In this paper, we introduce a Chinese understanding system, CUSAGA, which can understand technical Chinese texts and acquire knowledge from them. To facilitate the understanding process, CUSAGA can acquire a semantic grammar of the technical field, from which the Chinese texts arc sclected, in a semi-automatic way under the guidance of the user.

The overall structure of CUSAGA is shown in appendix A.

3 THE K<u>NOWLED</u>GE BASE <u>OF CUS</u>AGA

Due to the characteristics of Chinese, technical Chinese in particular, CUSAGA processes a Chinese text through different levels and divides the knowledge base into several parts.

The first part is the knowledge at morphology level. It contains a lexicon, the knowledge about how the Chinese characters constitute a word, and the context knowledge dealing with the ambiguity problem in segmenting words.

The second part is the knowledge at syntax level. It contains a semantic grammar, which is domain dependent and can be changed when the application domain is changed. At present, the semantic grammar being used is diagnosis oriented and has semantic classes like < person > , < time >, < feeling >, < symptom >, < disease >, and so on. It describes phrases and clauses in terms of these semantic categories.

The third part is the knowledge related to the deep structure of sentences. It contains some semantic patterns. The information in a semantic pattern shows which case slots can occur in this pattern and gives semantic restrictions for filling them.

* Supported by the TWAS grant GR 86 / 97.

The fourth part contains some domain-dependent knowledge. Technical Chinese allows ellipis of which the meaning can be figured out only if additional knowledge about the domain is provided. In addition, the required information is often not stated explicitely in these materials. Without the help of such knowledge, it is impossible to decide the referents for pronouns and abbreviations and to extract the required information.

1550 Speech and Natural Language

The partition of knowledge base makes the possible change of application domain more easy.

4 THE CHINESE UNDERSTANDING SUBSYSTEM IUC

IUC is supported by CALAS. The morphological knowledge and grammar learned by CALAS form the major part of the knowledge base of IUC. The following subsections discuss four separate phases of IUC in detail.

4.1 The Segmentation of Sentences

The segmentation of sentences is performed by the segmentation module in the following way: Any sentence, as a sequence of characters, is scanned from left to right. The segmentation information of the characters is used to recognize a sequence of reasonable words. If the segmentation information is not enough, then the lexicon will be used to look up words. If ambiguity occurs (i.e., more than one word is selected), then context dependent and domain dependent knowledge will be used. The segmentation of sentences in IUC is efficient, because multiple knowledge sources are used. For example, the Chinese sentence "左视乳头色泽深" can be segmented into a sequence of words: 左 (left), 视乳头 (optic), 色泽 (color), 涤 (dark). set up the meaning frame of the sentence. The derivation tree can be transformed into the meaning frame: dgr3 (Agent: c001; Location: PF(Direction: cb02; Eye: bk04)). Here, dgr3 is a predicate of the case frame, PF indicates a phrase frame.

4.4 Understanding Chinese Text

There are some problems when extracting the required information based on the meaning frames of the sentences. There may be certain pronouns and abbreviations in a Chinese text. Some slots of the meaning frames may have no actual values at all so that these frames can not express the meanings of the sentences clearly. In addition, much information needed by the user is often implied in a set of the meaning frames of the text and can not be discovered directly. This module is designed for solving these problems. The knowledge about the context and the application domain is applied in this process. For example, an information list can be got by extracting information based on the meaning frames of a patient record:

1) Personal State(sex)	: 1. male	2. semale	Answer:	1
2) Eye(direction):	1. left	2. right	Answer:	1
3) Optic(color):	1. light	2. moderate		
	3. dark		Answer:	3

Each time when a sentence is to be parsed, the lexicon is first searched for semantic parameters of all words in this sentence (each semantic parameter represents a category of words), then some heuristic information will be used to reduce the search space of appropriate rules. The rules are used in a recursive manner until a complete sentence structure is built up. top-down processing and depth-first search with backtracking is required for parsing.

The tree structure built by the parser not only describes the syntax structure of the sentence but also implies semantic information. This reduces the problem of ambiguity, makes the semantic analysis more efficient, and makes an efficient matching strategy for resolving ellipsis possible. For example, a derivation tree of the sentence mentioned above looks like follows: S(BZU (ff (cb02), b (bk04)), MC (c (c001), f (dgr3))). Here, the grammar rules (S::= BZU MC, BZU::= ff b, MC::= c f, ff::= {cb02...}, b::= {bk04...}, c::= {c001...}, f::= {dgr3...}) are used and cb02, bk04, c001, dgr3 are the semantic parameters of the words \pounds , 视乳 头, 伯泽 and 深.

4.3 Semantic Analysis

The semantic analysis is a recursive process for choosing appropriate components from the tree structure built by the parser and fitting them into the meaning frame of the sentence. This meaning frame is an improved version of Fillmore's case frame. The case slots in this frame are not filled directly with words of the sentence. Instead, they are filled with semantic parameters of these words to unify the representation of different phrases or sentences of the same meaning. Since predicates in Chinese are not only verbs, but also can be adjectives and nouns, we have set up different sorts of case frames for verbs, adjectives and nouns respectively. There are some special meaning frames: phrase frames, sentence frames and text frames which are not associated with a predicate. The semantic analysis begins at the top level of the tree structure of the sentence and then goes down to other levels of this tree in turn. With help of the semantic patterns in the knowledge base, cach level of the tree structure will be examined to look for predicates, then the semantic patterns of these predicates will be used to

5 THE LEARNING APPRENTICE CALAS

.....

Our experience with IUC showed that the bottleneck in the process of Chinese understanding using IUC is the building of a appropriate semantic grammar by the user himself. This work is very time consuming. Therefore we started the second subproject of CUSAGA, namely CALAS, to solve this problem by computer-aided grammatical inference. For natural languages, especially as complicated as Chinese, there exist a lot of serious difficulties at present time in designing fully automatic grammatical inference systems. So the computer-aided inference is a reasonable approach.

CALAS's final target is to learn a semantic grammar from a set of Chinese technical materials. The learning process is divided into two major stages. The first stage learns morphological knowledge and the second one learns the grammar itself.

Several researchers have achieved significant advances in this area. [Gold, 1967] and [SolomonofT, 1964] did some carly work. [Feldman, 1972] defined the grammatical inference problem and proved some decidability results about it. [Horning 1974] developed a procedure for grammatical inference. [Knobe, 1976] mentioned a method for inferring context-free grammars. Until recent years [Berwick, 1979, 1985] still works on this area. A good survey of inference research can be found in [Angluin and Smith, 1983]. [Bai, 1987] described a discoverer of Chinese grammar based on distribution analysis.

CALAS borrows many ideas from the results mentioned above. For example, its replacement operation is an improved version of the techniques used by Knobe. But CALAS distinguishes itself in many aspects, e.g., its learning process is incremental and can be performed under the guidance of an ordinary user without any special knowledge. Together with the increase of the knowledge base of CALAS, less and less guidance is needed from the user. CALAS will finally acquire the ability to distinguish between grammatical and ungrammatical Chinese sentences in the related fields.

5.1 Learning Morphological Knowledge

In CALAS the morphological knowledge is not provided a priori. It

Ruqian, Yinghui and Xiaobin 1551

is learned by CALAS during segmentation process with some help from users. The first sentence is always segmented by the user himself. This information serves as the initial morphological knowledge for the machine. Starting from the second sentence, the segmentation module does the work on its own by using this initial knowledge. If the segmentation is not reasonable (this is usually unavoidable, especially at the beginning of the learning process), the user corrects it. At this moment, through observing what the user does CALAS learns knowledge from the user's behavior. It extends the lexicon by adding into it some new words and acquires some new segmentation information. It can also acquire context dependent knowledge by observing how the user segments a sentence when ambiguity occurs. In this way, the segmentation module is getting more and more capable of segmenting a sentence.

5.2 <u>Learning Grammatical Knowledge</u>

Learning of a grammar is generalization based. The learning component consists of an informant (a teacher), a rule base (the grammar rules, RB for short) and a learning algorithm (for generalizing rules). The teacher introduces positive samples (legal sentences) to CALAS one at a time. CALAS enriches its RB while processing the input sentences. No negative samples are used. such that adding the production

(2)

to the grammar adds T to the language.

S:: = B1 B2---Bm

If D = D1 D2...Dn is a string of terminals and nonterminals and the sample sentence is derivable from D, then the production

S:: = D1 D2---Dn (3) will add the sample to the language. We shall call such a string a partial parse, and call (3) partial parse rule.

C Structural Ambiguity

Because of the complexity of Chinese, there are often more than one hierarchical structures corresponding to a same phrase or sentence. We call such situations <u>structural ambiguity</u>. For example, the simple sentence "网 膜 血 管 正常"(retina vascular normal), has at least two different hierarchical structures: (((retina) (vascular)) (normal)) and ((retina) ((vascular) (normal))).

D Key Words

Any sentence describes some objects or events. We define these objects or events as <u>key words</u> of the sentence. For instance, in the sentence "右 网膜 正常"(right retina normal), the word 网膜 is

CALAS generalizes grammar rules on the basis of its category base, current grammar and current sentence. A few number of categories are provided initially. The category base is then built step by step by adding new categories into it, learned from new input sentences.

The initial grammar can be empty, then the category strings corresponding to the word strings of the first input sentence will become the only rule of the initial RB. We'll assume that the teacher always knows whether or not a word string is grammatical but he needn't to know the concrete grammar rules.

The main problem of learning grammar rules is that the learning algorithm must be able to update the RB with the help of the user, when the current sample is not yet in the language the current grammar can generate. The updates might involve the modification and/or the deletion of existing rules, or the addition of new rules.

5.2.1 Basic Concepts of the Learning Algorithm

This section will discuss some formal concepts used by the learning algorithm.

A The Representation of Grammar Rules

The representation of grammar rules is based on the BNF notation. For instance, A::=A1 A2...An, is a rule, where A is a nonterminal, A1, A2,... An are n strings of terminals and (or pertorminals

the key word of the sentence.

E Basic Structures

We already noticed that a same sentence may have more than one structures, but some word strings appear often as a whole in the sentences and thus have the characterization of being 'grouped'. We introduce the concept of <u>basic</u> structures for such minimum 'grouped' characterizations. The kernel of a basic structure is its kernel word. We also call a basic structure whose kernel word is a key word, <u>a key phrase</u>. In general, the grammatical function of a key phrase as a whole is consistent with its kernel word.

For instance, in the sentence: "左 视乳头 色泽 深"(left optic color dark), the phrase, 左 视乳头, is a basic structure and its kernel is 视乳头.

F Matching Strings

We say that we have a match of two rules (one of which is a partial parse rule) if:

(1) they are in the form:

	X:: = A1 P1 Q1 R B1	(4)
and	X::= A2 P2 Q2 R B2	(5)

where X is a terminal symbol, A1, A2, P1, P2, Q1, Q2, R, B1 and B2 are strings of terminals and / or nonterminals, rule (4) is a partial parse rule.

(2) P1 and P2 are nonempty. P1 is the key word or the key phrase of the partial parse.

(3) (P1 = P2) or (P1 is the kernel of P2) or (P2 is the kernel of

and / or nonterminals.

B Partial Parse

Given any sentence T = A1 A2...An, it is trivial to find a production which will add T to the language, namely,

S::= A1 A2···An (1) where S is the start symbol of the grammar. But productions of this form may be much too specific. There may be some hierarchical structure in the A's. That is, there is probably a string B1 B2...Bm where some of the B's are nonterminals, P1).

(4) R is nonempty or R, B1 and B2 are all empty.

(5) Q1 and Q2 are nonempty. We shall refer to these corresponding nonempty strings as matching strings. And we shall say that Q1 and Q2 appear in <u>similar contexts</u> and call them a pair of similar strings.

Note that there may be several matches between two rules.

G The Main Strategies of the Inductive Learning

1552 Speech and Natural Language

The basic operation of replacement has two options:

(1) Disjunction: For a pair of matching strings, t1 and t2, which appear on right-hand sides of rules, build new rules D::=t1 and D::=t2 and add them to the grammar, where D is a nonterminal. Every occurrence of the strings t1 and t2 in existing rules is replaced by D.

Each time CALAS finds a pair of similar strings, it tries to build a common grammatical construct for these two strings. Note that only when t1 and t2 embody a common grammatical construct, can CALAS apply disjunction. The determination of key words and basic structures is just for finding as quickly as possible the matching strings embodying a common grammatical construct.

(2) Shorthand substitution: When a string, t, appears on right—hand sides of more than one rules, it is often good to create a new nonterminal, A, replacing all occurrences of t and add the rule A:: = t to the grammar.

Also note that only when t can be looked as a whole, can CALAS apply shorthand substitution.

5.2.2 Overview of the Learning Algorithm

A Abstract the Sample Sentences to Grammar Rules

Assume that CALAS has already processed a few samples, and has learned a partial grammar. Let the current input sentence be:

(6)

plete hierarchical structure of a sentence. But in fact, it is not easy to determine this structure, and even if it is possible, this structure is often not unique. In CALAS, the user only needs to point out basic structures in the sample rather than to provide the complete hierachical structure of it. So we can avoid structural ambiguity. The phrase, 左 视乳头, in (6) is a basic structure. So the corresponding rule produced by CALAS is

 $BZU::= ff \qquad b \qquad (13)$

We can note that (13) already exists in the rule base, thus it dose not modify the old rules.

Now the partial parse rule (12) is replaced by the rule

S::=BZU * c * f (14) where the kernel of the key phrase, BZU, is b.

Now the current construct of (6) is: (((lcft) (optic)) (color) (dark)).

C Inductive Learning on the Rule Base

After the basic structure is determined as above, CALAS performs match operation. It searches out matching strings, and suggests some grammatical structures to the user. Once they are confirmed by the user the rule learner can apply disjunction and shorthand substitution to generalize and simplify the RB.

C.1 Search for Matching Strings

左视乳头 色泽深

(left optic color dark)

which is already correctly segmented by the segmentation module.

Each word belongs to some categories. The first thing the algorithm must do is to determine the categories of all words in the current sentence.

We assume that CALAS already has the following categories in its category base:

b::={网膜(rctina),视乳头(optic),…}, f::={正常(normal),深(dark),…}, c::={色泽(color),…}, ff::={右(right),左(left),…}------

Then the string of categories corresponding to the input sentence is the following

ff b c f (7) Rewrite (7) in the form of a rule:

 $S::=ff \quad b \quad c \quad f \tag{8}$

Evidently (8) is a partial parse rule. For the moment, this rule is not added to RB. Assume that the following rules exist in RB, where each rule is produced by the sample phrase to the right of it.

1. $BZU::= ff b$	(右网膜)	(9)
2. $BZU::= b$	(网膜)	(10)
3. S:: = $BZU f$	(右网膜正常)	(11)

In fact, what the rule learner does is to find out matching strings embodying common grammatical constructs from the current partial parse rule and rules in RB, and then to replace them by nonterminals.

Since the match operation is based on key words and basic structures, CALAS must search them out before performing match operation. The search is done from right to left of the partial parse rule. Thus a string which can match 'f' in (14) is looked for in the current RB. But none of the rules in the RB contains a string which can match 'f'. Therefore CALAS tries to find out a string which can match 'c f' in (14). This time it does find such a string, namely 'f' in (11), which appears in a context similar to that of 'c f' in (14). Hence a pair of similar strings, 'f' and 'c f', is found.

C.2 Interactive Induction Strategies

Early in 1967, GOLD proved that if a program is only given an infinite sequence of positive samples the program can not determine a grammar for the context-free language within any finite time.

Fortunately, in most concrete situations, additional information (eg. an informant) is available which can help to limit the number of alternatives in the learning process so that the grammar of a reasonable language can be learned.

In CALAS, when the current partial grammar can not accept the current sample, CALAS will automatically create some new rules to try to accept it and produce some clauses to be used for checking whether these new rules are grammatical or not.

In our example, after having found the pair of similar strings, 'f' and 'c f', CALAS will produce word strings, pl = 正常 and p2 = 色泽深 and their relative contexts, cl = f网膜正常 and c2 = 左视乳头色泽深 respectively. If the

B <u>Determine Key Words and Basic Structures in</u> the Input Sentence

In (6), both of 视乳头 and 色泽 are key words. By using star * to indicate key words or key phrases, (8) is replaced by the follow-ing partial parse rule:

S::= ff b * c * f (12)

Some grammatical inference systems try to determine the com-

user agrees that both p1 and p2 are descriptions of the category b and are valid grammatical strings, then he can give a name, MC, as an indication of such a description. So the rule learner can automatically add the new rules 4: MC::=c f (色泽茶) (15) 5: MC::=f (正常) (16) to RB, replace both f in (11) and c f in (14) by MC, and

obtain rules

3':S:: = BZUMC(11')andS:: = BZU *MC(14')

Ruqian, Yinghui and Xiaobin 1553

It can be seen that the (11') can do all (14') can do, therefore (14') can be deleted. Now the current grammar can accept (6).

So far the construct of (6) is: (((left) (optic)) ((color) (dark))).

D Complete the Grammar After Processing the Sample

The grammar rules now look like as follows:

1: BZU::= ff b	(右网膜)
2: $BZU::= b$	(网膜)
3: S::= BZU MC	(c.g.,右网膜正常)
4: $MC::=c$ f	(色泽深)
5: MC::= f	(正常)

In another word, in the course of the grammatical inference and with the increase of samples the knowledge base is modified and enlarged.

6 THE IMPLEMENTATION OF CUSAGA

The system CUSAGA is implemented with about 12000 lines of PASCAL code and is running on the system UV68000.

Now, CUSAGA has been applied to understand patient records of glaucoma written in Chinese and to acquire knowledge for Inference. Information and Control, 7:1-22,224-254.

- [Feldman, 1972] Feldman J. Some Decidability Results on Grammatical Inference and Complexity. Information and Control, 20:244-262
- [Horning, 1971] Horning J. J. A Procedure for Grammatical Inference. In Proc. IFIP Congress, Ljubljana.
- [Knobc and Knobc, 1976] Knobe B. and Knobc K. A Method for Inferring Context-free Grammars. <u>Information</u> and Control, 31:129-146.
- [Berwick, 1979] Berwick R. Learning Structural Descriptions of Grammar Rule from Example. In Proceedings of the Sixth International Conference on <u>Artificial Intellig</u>ence.
- [Berwick, 1985] Berwick R. <u>The</u> Acquisition of Syntactic Knowledge. The MIT Press, 1985. ~~~~

[Angluin, 1983] D. Angluin and C. Smith. Inductive Inference: Theory and Methods. <u>ACM Computing Surveys</u>, 15:237-269.

- [Tianshun, 1985] Yao Tianshun. A Chinese query language for data base interfaces. T.R. Northeastern Institute of Technology.
- [Zao, 1987] Zao Zhcnxi. Knowledge-based hierarchical Chinese understanding and its knowledge representation. T.R. China University of Science and Technology.
- [Xic, 1988] Xic Shiming. A hybrid Chinese-English language understanding and generation system. M.S. thesis, Institute of Mathematics, Academia Sinica.

[Guo, 1985]Guo Yuanxin. A Recursive Model for Chinese Syntax.

constructing the knowledge base of an expert system in glaucoma. We have made two experiments. In the first one, the semantic grammar used by IUC was summarized by one of the authors after reviewing a set of patient records. In the second one, it was acquired with the help of CALAS.

The following data illustrate the results of the two experiments, where 'n umber of patient records' is called nprc for short.

	npre	increase	number-of-rules	increase
	1	1	26	26
	5	4	61	35
1st	15	10	94	33
	25	10	110	16
	40	15	119	9
	1	1	26	26
	3	2	79	53
2nd	5	2	93	14
	7	2	101	8
	8	1	103	2
	43	35	116	13

Note that CALAS acquires nearly 90% of the rules (103) with only less than 1/5 of the records (8), while the human-being needs to review about 5/8 of the records (25) in order to get to the same level (110). This shows the rapid convergence of the semantic grammar acquired by CALAS. Another important advantage of CALAS is the saving of time to review the records by human-being.

We must point out that the approaches of CALAS are not

M.S.thesis, Qinghua University.

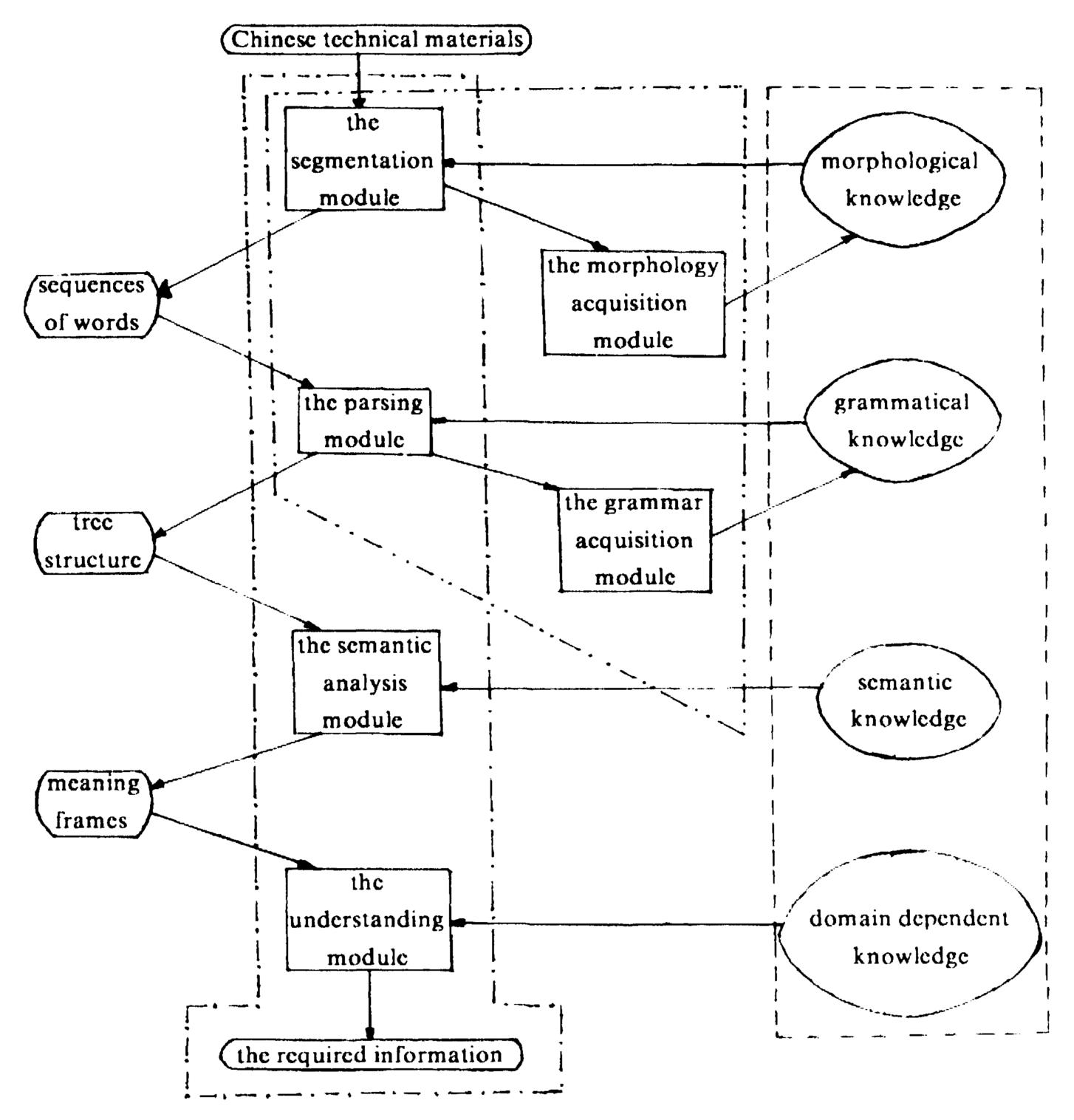
- [Wang, 19831 Wang Xilong. A System for Understanding Chinese. In proceedings of the international conference on Chinese information processing.
- [Wang, 1987] Wang Yeyu. A Mathematical Model for Chinese Natural Language. M.S. thesis, Institute of Mathematics, Academia Sinica.
- [Bai, 1985] Bai Shuo. Computer-aided Discover of Chinese Syntax. M.S thesis, Department of Computer Science, Beijing University.

limited to acquiring semantic grammar, but can also be used for other grammar forms. Nor is CUSAGA limited to be used in certain technical field, by changing the background knowledge base, it can be switched to other domains easily.

REFERENCE

[Gold, 1967]M.Gold. Language Identification in the Limit. Information and Contral, 10:p447. [Solomonoff, 1964] Solomonoff, R. A Formal Theory of Inductive

1554 Speech and Natural Language



Legend:

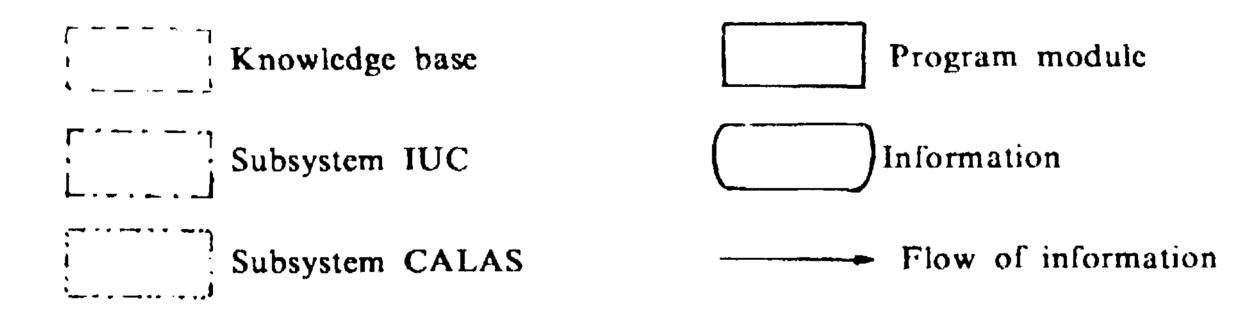


Figure 1

Ruqian, Yinghui and Xiaobin 1555