NEC:
DESCRIPTION OF THE VENIEX SYSTEM AS USED FOR MUC-5

Kazunori MURAKI, Shinichi DOI and Shinichi ANDO
NEC Corp. Information Technology Research Laboratories
Human Language Research Laboratory
4–1–1, Miyazaki, Miyamae-ku, Kawasaki 216, JAPAN
E-mail: {k-muraki, doi, ando}@hum.cl.nec.co.jp
Phone: +81–44–856–2148, Fax: +81–44–856–2238

INTRODUCTION

NEC Corporation has had years of experience in natural language processing and machine translation[1, 2, 3, 4, 5], and currently markets commercial natural language processing systems. Utilizing dictionaries and parsing engines we have already had, we have developed the VENIEX System (VENus for Information EXtraction) as used for MUC-5 in only three months. Our method is to apply both domain-specific keyword-based analysis and full sentential parsing with general grammar[6, 7]. The keyword dictionary of VENIEX contains about thirty thousand entries, whose semantic structures are sub_ME_Capability frame, and the parsing and discourse processing are controlled with the information given in this semantic structure of keywords. The resulting scores of VENIEX for formal run texts were from 0.7181(minimum) to 0.7548(maximum) in Richness-Normalized Error and 48.33 in F-MEASURES(P&R).

SYSTEM ARCHITECTURE

The overall system architecture is shown in Fig. 1. An input text is divided into sentences and each sentence is processed separately. ME_Capability frames are extracted from each sentence. An example of the procedure of information extraction from one sentence by VENIEX is shown in Fig. 2-4.

The characteristic modules of VENIEX are as follows:

- Keyword Dictionary which contains about thirty thousand entries, whose semantic structures are sub_ME_Capability frame,
- Parser which generates ME_Capability frames by correlating keywords during full sentential parsing, whose process is controlled with the information in this semantic structure of keywords,
- Discourse Processor which combines ME_Capability frames of each sentence.

We call this lexical-information-driven method for parsing and discourse processing "Lexical-Discourse-Parsing". This method utilizes the merits of both domain-specific keyword-based analysis and full sentential parsing and discourse processing with general grammar. It also reduces expenses of general parsing and discourse processing.

Preprocessor

This module divides an input text into a header and a body of the text, and stores the document number, date and source information from the header for entry into the template. It also divides the body of the text into sentences, which will be processed separately during morphological analysis and parsing.

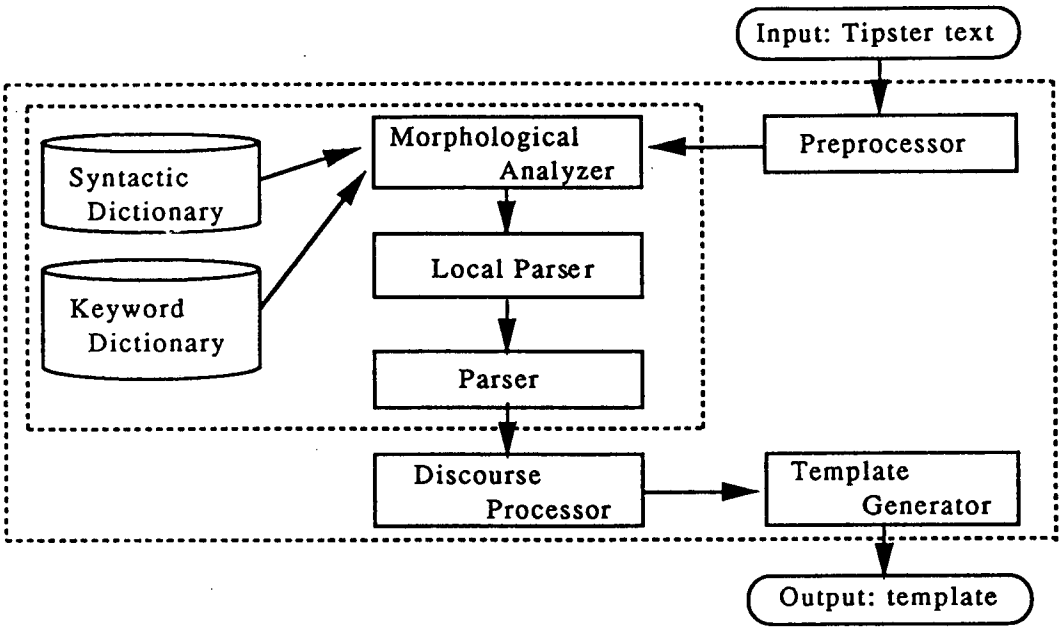


Figure 1: System Architecture of VENIEX

Dictionaries

Our system utilizes two dictionaries, a syntactic dictionary and a keyword dictionary. Both dictionaries are converted from the machine translation dictionaries we had developed. The syntactic dictionary contains about ninety thousand entries and the keyword dictionary contains about thirty thousand entries, including the names of corporations, pieces of equipment, devices, place names, etc. Also, we extracted the names we didn't have in our original dictionaries from the Tipster corpus, and enlarged the keyword dictionary.

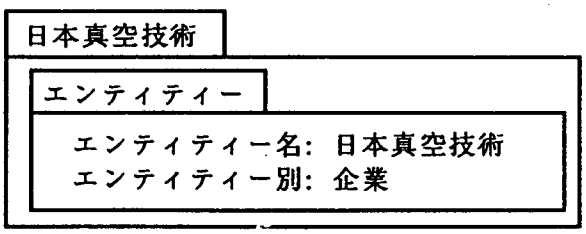
We added semantic structures which are sub_ME_Capability frame –partial structure of ME_Capability frame– to the entries of the keyword dictionary. Examples of the sub-frames are shown in Fig. 2.

Fig.2-a) is an example of an Entity sub-frame, which provides slots for a name and a type of an entity. This sub-frame can provide other-name-slot whose value is a list of other names of the entity including nicknames and abbreviations, such as "NEC" and "日電" of "日本電気".

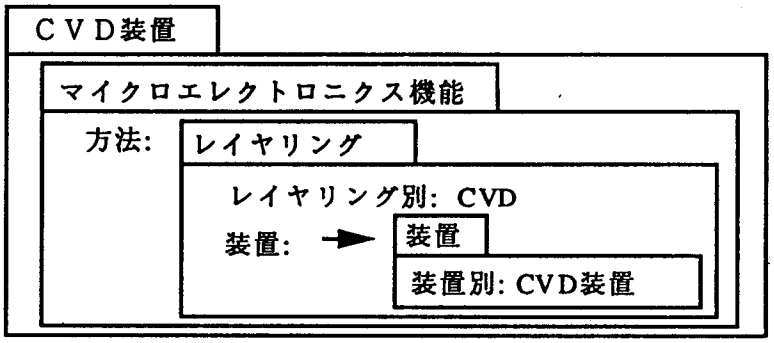
Fig.2-b) is an example of a ME_Capability sub-frame, which provides slots for the process type and detailed information of the process including its type and the equipment used. The words to which a ME_Capability sub-frame is added are extracted from technical term dictionaries of microelectronics and the Tipster corpus.

Fig.2-c) is an example of a Relation sub-frame, which is added to words representing the relation between words with an Entity sub-frame and words with a ME_Capability sub-frame. These sorts of words are generally Japanese verbs. The Relation sub-frame provides case slots with a case marker –Japanese postpositional particles– representing grammatical relations. Each case slot contains a sub-slot representing whether the filler of this case slot is a word with an Entity sub-frame or a word with a ME_Capability sub-frame. If it

a) Entity
ex)



b) ME_Capability
ex)



c) Relation
ex)

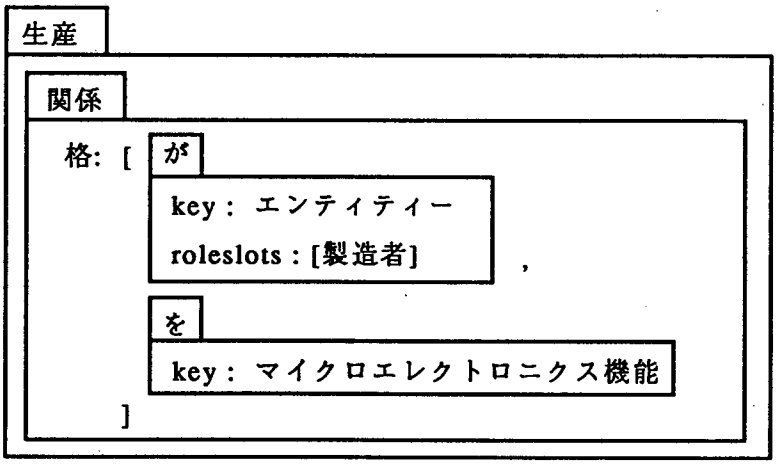


Figure 2: Example of sub_ME_Capability Frame

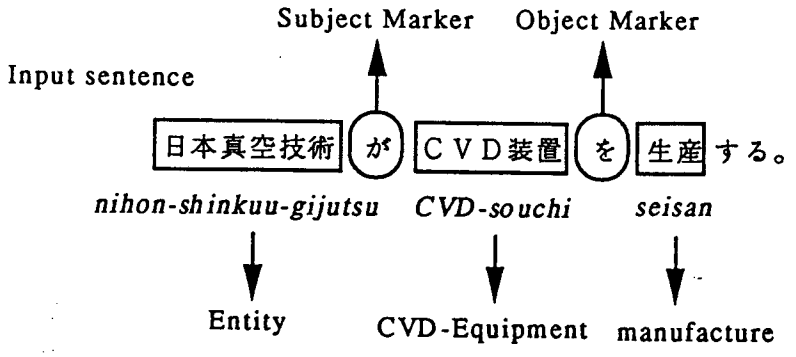


Figure 3: Example of Input Sentence

is a word with an Entity sub-frame, the case slot also contains a sub-slot representing the role of the Entity sub-frame to a ME_Capability sub-frame, whose value is a list of "開発者 (developer)", "製造者 (manufacturer)", "配給者 (distributor)" or "購入 / 利用者 (purchaser_or_user)". Therefore, the Relation sub-frame in Fig.2-c) means that:

- The Japanese verb "生産 (manufacture)" has two case slots.
- The filler of the first slot with a subject marker "が" is a word with an Entity sub-frame, whose role to a ME_Capability sub-frame is "製造者 (manufacturer)".
- The filler of the second slot with an object marker "を" is a word with a ME_Capability sub-frame.

Morphological Analyzer

This module divides input sentences into morphemes and gives every morpheme lexical attributes with a syntactic dictionary and a keyword dictionary. For example, an input sentence "日本真空技術がCVD装置を生産する。(Nihon–shinkuu–gijutsu manufactures a piece of CVD-equipment.)" is divided as shown in Fig. 3, and the semantic structures shown in Fig. 2 are given to morphemes "日本真空技術", "CVD装置" and "生産".

If a morpheme is encountered that doesn't exist in the dictionaries, it is marked as an unknown word and its part of speech is estimated from neighboring morphemes. For example, in a text that contains many nouns, the recognition of unknown words becomes an important function because these words may be important proper nouns. Numerical values are also tagged with the same kinds of information as words because they often perform as content words, and are often useful for determining sentence structures.

Local Parser

This module re-collects morphemes given by the Morphological Analyzer and produces phrases. It also combines the ME_Capability sub-frames given to the words in a phrase, and assigns a new combined ME_Capability sub-frame to the output phrase.

This module also deduces keywords from particular suffixes and patterns. For example, the nouns preceding the suffix "社" or "会社" is considered as business entities. Unknown noun preceding parentheses inserted a place name can be business entities, too.

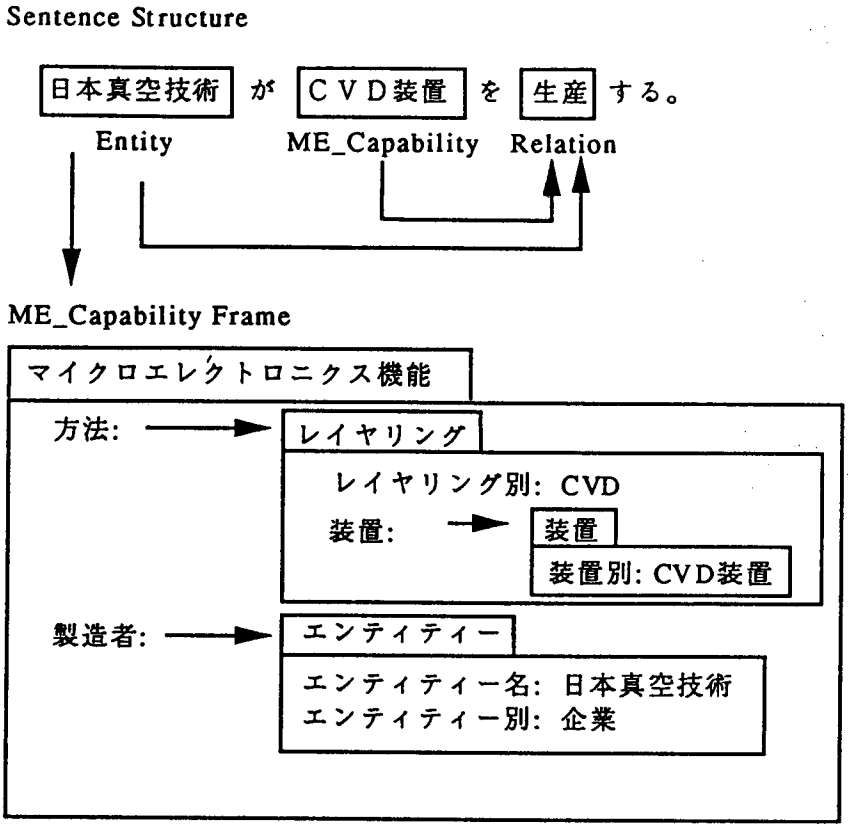


Figure 4: Example of Information Extraction

Parser

This module re-collects phrases produced by Local Parser and outputs parse trees and semantic structures, which are ME_Capability frames. Its function involves not only parsing but also semantic interpretation, lexical disambiguation and information extraction. The main body of the analyzer is a unification-based chart parser, and the parsing strategy is bottom-up breadth-first. The solution with the highest preference score is selected. Our Local Parser and Discourse Processor are based on the same parsing engine and differ only in parsing rules. Sharing engines and functions by modules, we can efficiently develop the VENIEX system.

The Parser can handle a wide variety of complex sentences. It analyzes and generates modifying connections between phrases and the relation between keywords. It constructs semantic structures which are ME_Capability frames from sub-frames described in a keyword dictionary by correlating keywords during full sentential parsing, whose process is controlled with the information in the sub-frames. For example, as illustrated in Fig. 4, the Parser recognizes the structure of the sentence "日本真空技術が CVD 装置を生産する。" and constructs a ME_Capability frame from sub-frames shown in Fig. 2.

This module can also deduce keywords. If an unknown noun fills a Relation sub-frame's case slot whose filler must be a word with an Entity sub-frame, this noun can be considered as an entity.

In addition, the Parser recognizes special expressions whose sub_ME_Capability frames are used for discourse processing. It selects the most important Entity sub-frame and the ME_Capability sub-frame, and also analyzes the Entity sub-frame and the ME_Capability sub-frame represented by anaphoric expressions.

Parser keep these sub-frames respectively in "currentEnt" slot, "currentME" slot, "anaphorEnt" slot and "anphorME" slot. We will later show examples of these slots with a walkthrough example.

Though it is not illustrated in Fig. 1, VENIEX has another module as a fail-safe system between the Parser and the Discourse Processor. If the Parser cannot analyze an input sentence and outputs only fragments of ME_Capability frame, this Postparser module re-collects and combines the fragments without considering the sentence structure.

Discourse Processor

This module combines ME_Capability frames generated by the Parser into frames representing content of the whole article. It recognizes relation among the ME_Capability frames by resolving co-reference for entities and microelectronics. The co-reference resolution is achieved by unifying "currentEnt" with "anaphorEnt" and unifying "currentME" with "anphorME". VENIEX can resolve co-reference represented by a wide variety of expressions: anaphoric expression (identical and unidentical), cleft sentence, ellipsis, name of Entities, etc[7]. We will later show an example of this process with a walkthrough example as well.

Template Generator

In VENIEX, the outputs of the Discourse Processor are ME_Capability frames. In other words, essential information has already been extracted during morphological, syntactic and discourse analysis. All that remains is to transform the frames and the information of the input article stored by the Preprocessor to the output templates in the official form.

PROCESSING WALKTHROUGH TEXT

Overview

VENIEX has two steps for ME information extraction; 1) extracting ME_Capability frames separately from each sentence, 2) combining the frames above into frames representing content of the whole text. This method has two tasks in constructing a body of knowledge with small pieces of information contained in more than one sentence. First, it must construct new information with pieces of partial information scattered in different sentences. Second, it must identify identical information represented by different expressions. VENIEX attains these tasks by discourse processing on surface expressions focused on ellipsis, anaphora and so on. The walkthrough text, however, has discourse problems that can't be solved with that particular surface process. Therefore VENIEX can't merge the information sufficiently and outputs two ME objects for only one ME object in the text. Also, VENIEX fails in complement of ellipsis and extracts only one entity for two entities. As a result, the evaluation of walkthrough text is 66.67 P&R.

Morphological Analyzer

The Morphological Analyzer divides a sentence into morphemes and assigns corresponding syntactic information to each morpheme using the syntactic dictionary. At the same time, it assigns some information from the keyword dictionary to morphemes.

Fig. 5 below shows the result of morphological analysis of the 1st sentence of the walkthrough.

```
/半導体/製造
   {key 関係{slot [体言句{csh が; ckey エンティティー; rolslots [製造者]},
                   体言句{csh を; ckey マイクロエレクトロニクス機能}]}}
/装置
   {key マイクロエレクトロニクス機能{方法 top{装置 装置}}}
/などの/大手
   {key 関係{slot [体言句{csh が; ckey エンティティー; rolslots [開発者, 製造者, 配給者]},
```

```
                    体言句{csh を; ckey マイクロエレクトロニクス機能}]}}
/メーカー/、/日本真空技術
   {key エンティティー{エンティティー別 企業; spell 日本真空技術}}
/(/本社/神奈川県茅ケ崎
   {key 地名{gazette 日本(国) 神奈川(県) 茅ケ崎(市); level 3}}
/市/、/社長/高村/甚平/氏/)/は/米国
   {key 地名{gazette 米国(国); level 1}}
/の/半導体/製造
   {key 関係{slot [体言句{csh が; ckey エンティティー; rolslots [製造者]},
                    体言句{csh を; ckey マイクロエレクトロニクス機能}]}}
/用/拡散/炉/の/最大手
   {key 関係{slot [体言句{csh が; ckey エンティティー; rolslots [開発者, 製造者, 配給者]},
                    体言句{csh を; ckey マイクロエレクトロニクス機能}]}}
/、/BTUインターナショナル
   {key エンティティー{エンティティー別 企業;
                       別称 [ビーティーユーインターナショナル, ビーティユーインターナショナル];
                       spell BTUインターナショナル}}
/社/(/本社/マサチューセッツ州
   {key 地名{gazette 米国(国) マセチューセッツ(県); level 2}}
/)/と/米国
   {key 地名{gazette 米国(国); level 1}}
/に/合弁/会社/を/設立/、/半導体/製造
   {key 関係{slot [体言句{csh が; ckey エンティティー; rolslots [製造者]},
                    体言句{csh を; ckey マイクロエレクトロニクス機能}]}}
/の/重要/装置
   {key マイクロエレクトロニクス機能{方法 top{装置 装置}}}
/の/ひとつ/で/ある/金属膜用CVD(化学的気相成長法)装置
   {key マイクロエレクトロニクス機能{方法 レイヤリング{膜別 @@CVD; フイルム 金属;
                                                    装置 装置{装置別 @@CVD装置}}}}
/を/生産・販売
   {key 関係{slot [体言句{csh が; ckey エンティティー; roleslots [製造者, 配給者]},
                    体言句{csh を; ckey マイクロエレクトロニクス機能},
                    体言句{csh or(に, へ, へと);
                           ckey エンティティー; roleslots [購入者/利用者]}]}}
/する/こと/で/合意
   {key 補助用言{slot [体言句{csh が; ckey エンティティー},
                       体言句{csh で; ckey 関係}]}
/し/た/。/
```

Figure 5: Walkthrough — The result of morphological analysis —

The notation "/" in Fig. 5 is a delimiter of two morphemes. A morpheme recognized as a keyword is followed by a corresponding sub_ME_Capability frame, which is a partial structure of ME_Capability frame, loaded from the keyword dictionary. For example the word "製造", which means "manufacturing", has information that the entity which appears as the subject plays a manufacturer part of the object, the ME_Capability frame. The word "日本真空技術", which is a company name, has information that the type is company. The word "金属膜用CVD(化学的気相成長法)装置", which means "CVD equipment", conveys information that the type is layering and that the film is metal, and implies the existence of equipment.

VENIEX gathers the sub_ME_Capability frames and combines them into the ME_Capability frames.

Local Parser

The Local Parser recognizes a Japanese phrase by utilizing local patterns and the syntactic information given by the Morphological Analyzer. The Local Parser combines sub_ME_Capability frames in one phrase. The way of combination differs according to the sort of keywords; "エンティティー" (entity), "マイクロエ

レクトロニクス機能" (microelectronics), "関係" (relationship) and so on.

The output of the Morphological Analyzer is shown in Fig. 6.

```
/半導体製造装置などの
   {key マイクロエレクトロニクス機能{方法 top{装置 装置}}}
/大手メーカー、
   {key 関係{slot [体言句{csh が; ckey エンティティー;
                          roleslots [開発者, 製造者, 配給者]},
                   体言句{csh で;
                          ckey マイクロエレクトロニクス機能}]}}
/日本真空技術（本社神奈川県茅ケ崎市、社長高村甚平氏）は
   {key エンティティー{場所 日本（国）神奈川（県）茅ケ崎（市）;
                       spell 日本真空技術;
                       エンティティー別 企業}}
/米国の
   {key 地名{gazette 米国（国）; level 1}}
/半導体製造用
   {key 関係{slot [体言句{csh が; ckey エンティティー; roleslots [製造者]},
                   体言句{csh を; ckey マイクロエレクトロニクス機能}]}}
/拡散炉の
/最大手、
   {key 関係{slot [体言句{csh が; ckey エンティティー;
                          roleslots [開発者, 製造者, 配給者]},
                   体言句{csh で; ckey 関係}]}}
/ＢＴＵインターナショナル社（本社マサチューセッツ州）と
   {key エンティティー{場所 米国（国）マセチューセッツ（県）;
                       spell ＢＴＵインターナショナル;
                       別称 [ビーティーユーインターナショナル, ビーティユーインターナショナル];
                       エンティティー別 企業}}
/米国に
   {key 地名{gazette 米国（国）; level 1}}
/合弁会社を/設立、/半導体製造の
   {key 関係{slot [体言句{csh が; ckey エンティティー; roleslots [製造者]},
                   体言句{csh を; ckey マイクロエレクトロニクス機能}]}}
/重要装置の
   {key マイクロエレクトロニクス機能{方法 top{装置 装置}}}
/ひとつ/である/金属膜用ＣＶＤ（化学的気相成長法）装置を
   {key マイクロエレクトロニクス機能{方法 レイヤリング{膜別 @@CVD; フイルム 金属;
                                                        装置 装置{装置別 @@CVD 装置}}}}
/生産・販売する
   {key 関係{slot [体言句{csh が; ckey エンティティー; roleslots [製造者, 配給者]},
                   体言句{csh を; ckey マイクロエレクトロニクス機能},
                   体言句{csh or(に, へ, へと);
                          ckey エンティティー; roleslots [購入者/利用者]}]}}
/ことで/合意した。
   {key 補助用言{slot [体言句{csh が; ckey エンティティー},
                       体言句{csh で; ckey 関係}]}}
```

Figure 6: Walkthrough — The result of local parsing —

In Fig. 6, the entity "日本真空技術" acquires the new information by extracting the keyword which shows its location in the identical phrase.

Parser

The Parser recognizes the syntactic structure of each sentence in the input text. An ME_Capability frame in a phrase is combined with corresponding ME_Capability frames in other phrases if they have syntactic

relations. The way to combine the frames depends on the sort of each of the keywords.

The output of the Parser to walkthrough text is shown in Fig. 7. In Fig. 7, the number following a notation of "_" is an index. If two objects have a same index, these objects are the identical.

```
+++ 00 ++++++++++++++++++++++++++++++++

{entities [エンティティー{場所 米国 (国) マセチューセッツ (県) ; spell BTUインターナショナル;
                          別称 [ビーティーユーインターナショナル, ビーティユーインターナショナル];
                          エンティティー別 企業} _89131362,
           エンティティー{エンティティー別 企業; spell 日本真空技術;
                          場所 日本 (国) 神奈川 (県) 茅ケ崎 (市) } _89129292];
 total [マイクロエレクトロニクス機能{方法 top{装置 装置{製造者 □}};
                                  開発者 □; 製造者 □; 配給者 □; 購入者/利用者 □},
        マイクロエレクトロニクス機能{開発者 □; 購入者/利用者 □;
                                  方法 レイヤリング{フイルム 金属; 膜別 @@CVD;
                                                    装置 装置{装置別 @@CVD 装置;
                                                              製造者 [エンティティー _89129292]}};
                                  製造者 [エンティティー _89129292];
                                  配給者 [エンティティー _89129292]} _89129293,
        マイクロエレクトロニクス機能{方法 top{装置 装置{製造者 □}};
                                  購入者/利用者 □; 配給者 □; 製造者 □; 開発者 □}];
 currentEnt エンティティー _89129292;
 currentME マイクロエレクトロニクス機能 _89129293}

+++ 01 ++++++++++++++++++++++++++++++++

{entities [エンティティー{spell 日本真空技術; エンティティー別 企業} _92602607];
 total [マイクロエレクトロニクス機能{購入者/利用者 □; 開発者 □;
                                  方法 top{装置 装置{製造者 [エンティティー _92602607]}};
                                  製造者 [エンティティー _92602607];
                                  配給者 [エンティティー _92602607]} _92602606];
 anaphorME マイクロエレクトロニクス機能 _92602606;
 currentEnt エンティティー _92602607;
 currentME マイクロエレクトロニクス機能 _92602606}

+++ 02 ++++++++++++++++++++++++++++++++

{entities [エンティティー{エンティティー別 企業;
                          別称 [ビーティーユーアルバック, ビーティユーアルバック];
                          spell BTUアルバック} _96207855];
 anaphorEnt エンティティー _96207096}

+++ 03 ++++++++++++++++++++++++++++++++
{}
+++ 04 ++++++++++++++++++++++++++++++++
{}
+++ 05 ++++++++++++++++++++++++++++++++

{entities [エンティティー{ エンティティー別 企業; spell 日本真空技術} _102793590];
 total [マイクロエレクトロニクス機能{開発者 □; 購入者/利用者 □;
                                  方法 レイヤリング{膜別 @@CVD; フイルム 金属;
                                                    装置 装置{製造者 [エンティティー
_102793590]}};
                                  製造者 [エンティティー _102793590];
                                  配給者 [エンティティー _102793416,
                                          エンティティー _102793590]} _102793419];
 anaphorEnt エンティティー _102793416;
 currentME マイクロエレクトロニクス機能 _102793419}

+++ 06 ++++++++++++++++++++++++++++++++
```

```
{anaphorEnt エンティティー _108363917;
 anaphorME マイクロエレクトロニクス機能{購入者／利用者 □; 配給者 □; 開発者 □;
                                    製造者 [エンティティー _108363917];
                                    方法 top{ 装置 装置{ 製造者 [エンティティー _108363917]}}
                                   } _108363918;
 currentEnt エンティティー _108363917}
```

Figure 7: Walkthrough — The result of parsing —

In the 1st sentence (No. 00), for example, manufacture and distribution on the CVD equipment is extracted. The sentence consists of two simple sentences and the extracted information lies in the 2nd simple sentence. VENIEX recognizes that these two simple sentences share the nominative case, and combines the ME_Capability frames according to the path of information;

{"日本真空技術" - "合意した。" - {"生産・販売" - "ＣＶＤ装置"}}

({Entity - "agree" - {"manufacture and distribute" - Equipment}}).

This results in VENIEX extracting the ME_Capability frame, as shown in Fig. 7.

The Parser in VENIEX extracts 5 ME_Capability frames from 4 sentences; the 1st sentence, the 2nd sentence (No. 01), the 6th sentence (No. 05) and the 7th sentence (No. 06). As for the 6th sentence, VENIEX succeed in extracting 2 ME_Capability frames from the noun phrase, "日本真空技術が国内で生産・販売している金属膜用CVD装置", and an entire of sentence.

Meanwhile, the Parser keeps a sub_ME_Capability frame, which appears as an entity or microelectronics in the sentence, respectively in "entities" slot and "currentME" slot. Additionally, for an entity (sub_ME_Capability frame), which is the subjective case in the given sentence, the Parser keeps it in "currentEnt" slot. For a sub_ME_Capability frame represented by anaphoric expressions, the Parser also instantiates "anaphorEnt" slot or "anaphorME"slot. After anaphora resolution, it puts referred "currentEnt" slot or "currentME" slot in corresponding "anaphorEnt" slot or "anaphorME"slot.

Discourse Processor

The Discourse Processor merges ME_Capability frames the Parser output by utilizing "entities" slot, "currentEnt" slot, "currentME" slot, "anaphorEnt" slot and "anaphorME" slot. For example, the 2nd sentence in the walkthrough carries information that the entity "日本真空技術" distributes some equipment and the equipment which appears in anaphoric expression "同装置" refers to the CVD equipment in the 1st sentence. In processing the 2nd sentence, the Discourse Processor recognizes the expression "同装置" as an anaphoric expression and instantiates the "anaphorME" while extracting sub_ME_Capability frame (see Fig. 7). It checks consistency between the "anaphorME" slot and the "currentME" slot instantiated in processing the previous sentence and identifies these as the same object.

VENIEX makes two mistakes in discourse processing for walkthrough text.

One appears in ellipsis processing. Ellipsis of nominative in the 6th sentence must be resolved for extracting the entity which is the distributor. The distributor entity must be "BTU アルバック" in the 3rd sentence because it is clear, according to context, that the 2nd paragraph is written about its activities. But VENIEX selects "currentEnt" slot which is the nominative of the 2nd sentence, because it lacks knowledge to process a paragraph or joint venture.

The other mistake is caused by failure in merging ME_Capability frame in the 1st paragraph with one in the 2nd paragraph. These frames must be identical objects because the topic of the article is a joint venture, and a joint venture distributes often products of parent company. (We think, however, that the equivalence of the CVD equipment cannot be decided based only on these clues, and it is possible to interpret that this equipment are different.) VENIEX processes all ME objects separately when there is no specified referential

expression.

As a result, VENIEX output the template shown in Fig. 8.

```
<テンプレート-000452-1> :=
    記事符号: 000452
    発行年月日: 890804
    ニュース出所: "日経新聞 本紙朝刊"
    内容: <マイクロエレクトロニクス機能-000452-1>
    <マイクロエレクトロニクス機能-000452-2>
    完了年月日: 930820
    抽出時間: 2
<マイクロエレクトロニクス機能-000452-1> :=
    方法: <レイヤリング-000452-1>
    製造者: <エンティティー-000452-1>
    配給者: <エンティティー-000452-1>
<マイクロエレクトロニクス機能-000452-2> :=
    方法: <レイヤリング-000452-2>
    製造者: <エンティティー-000452-1>
    配給者: <エンティティー-000452-1>
<エンティティー-000452-1> :=
    エンティティー名: 日本真空技術
    場所: 日本 (国) 神奈川 (県) 茅ケ崎 (市)
    エンティティー別: 企業
<レイヤリング-000452-1> :=
    膜別: CVD
    フイルム: 金属
<レイヤリング-000452-2> :=
    膜別: CVD
    フイルム: 金属
    装置: <装置-000452-1>
<装置-000452-1> :=
    製造者: <エンティティー-000452-1>
    装置別: CVD装置
    状況: 利用中
```

Figure 8: Walkthrough — The template —

RESULTS AND FUTURE WORK

The resulting scores of VENIEX at formal run were from 0.7476(minimum) to 0.7858(maximum) in Richness-Normalized Error and 47.41 in F-MEASURES(P&R), which are shown in Table 1. We have improved the system a little after the formal run –only by debugging parsing rules, not by adding new rules and/or dictionaries–, and the current scores of VENIEX for formal run texts are from 0.7181(minimum) to 0.7548(maximum) in Richness-Normalized Error and 48.33 in F-MEASURES(P&R), which are shown in Table 2. The current scores for dry run texts are also shown in Table 3.

Though we have developed the VENIEX System in only three months, there wasn't so much difference in scores with other systems in MUC-5. But the scores were lower than we had expected. The main reason is the lowness of recall rate. We didn't have enough time to collect keywords, especially verbs representing the relations between entities and microelectronics.

We have developed many new functions for the MUC-5 system, such as co-reference resolution and keyword deduction. We have been evaluating these functions separately to judge weather they worked as we designed. For example, to evaluate the performance of keyword deduction function in the Local Parser and the Parser, we made an information extraction experiment without dictionaries of entities. The resulting

a) From the error-based score reports					
				Richness-Nomalized Error	
ERR	UND	OVG	SUB	Min-err	Max-err
67	55	30	14	0.7476	0.7858
b) From the recall-precision-based score reports					
		REC	PRE	P&R (F-Meseasure)	
All-Object		39	61	47.41	
Text-Filtering		66	83	——	

Table 1: Summary of our MUC-5 Score

a) From the error-based score reports					
				Richness-Nomalized Error	
ERR	UND	OVG	SUB	Min-err	Max-err
66	55	26	14	0.7181	0.7548
b) From the recall-precision-based score reports					
		REC	PRE	P&R (F-Meseasure)	
All-Object		39	64	48.33	
Text-Filtering		68	85	——	

Table 2: Current Scores for Formal Run Texts

a) From the error-based score reports					
				Richness-Nomalized Error	
ERR	UND	OVG	SUB	Min-err	Max-err
53	39	21	10	0.5566	0.5963
b) From the recall-precision-based score reports					
		REC	PRE	P&R (F-Meseasure)	
All-Object		55	71	61.91	
Text-Filtering		88	91	——	

Table 3: Current Scores for Dry Run Texts

a) From the error-based score reports					
				Richness-Nomalized Error	
ERR	UND	OVG	SUB	Min-err	Max-err
73	65	25	14	0.7639	0.8029
b) From the recall-precision-based score reports					
		REC	PRE	P&R (F-Meseasure)	
All-Object		30	64	40.87	
Text-Filtering		59	85	——	

Table 4: Scores of Experiment without Entity Dictionary

scores for formal run text are shown in Table 4. The result says that this function works well. Through the development of VENIEX system for MUC-5, we have learned that we can realize information extraction system with our natural language processing techniques. But to improve the system, we must make more detailed evaluation of performance of each function.

One of the biggest theme of future work is automated or semi-automated training of the system. We plan to develop a bootstrapping method to improve the system with iterating cycles of "refining system"–"evaluating the performance".

References

[1] Muraki, K., "VENUS: Two-phase Machine Translation System", *Future Generations Computer Systems*, 2, 1986

[2] Ichiyama, S., "Multi-lingual Machine Translation System", *Office Equipment and Products*, 18-131, August 1989

[3] Okumura, A., Muraki, K. and Akamine, S., "Multi-lingual Sentence Generation from the PIVOT interlingua", *Proceedings of MT SUMMIT III*, July 1991

[4] Doi, S., Muraki, K., Kamei, S. and Yamabana, K., "Long Sentence Analysis by Domain-Specific Pattern Grammar", *Proceedings of EACL 93*, April 1993

[5] Yamabana, K., Kamei, S. and Muraki, K., "On Representation of Preference Scores", *Proceedings of TMI-93*, July 1993

[6] Ando, S., Doi, S. and Muraki, K., "Information Extraction System based on Keywords and Text Structure", *Proceedings of the 47th Annual Conference of IPSJ*, October 1993 (in Japanese)

[7] Doi, S., Ando, S. and Muraki, K., "Context Analysis in Information Extraction System based on Keywords and Text Structure", *Proceedings of the 47th Annual Conference of IPSJ*, October 1993 (in Japanese)