A Case Analysis Method Cooperating with ATNG and Its Application to Machine Translation

Hitoshi IIDA, Kentaro OGURA and Hirosato NOMURA

Musashino Electrical Communication Laboratory, N.T.T.
Musashino-shi, Tokyo, 180, Japan

Abstract

This paper present a new method for parsing English sentences. The parser called LUTE-EJ parser is combined with case analysis and ATNG-based analysis. LUTE-EJ parser has two interesting mechanical characteristics. One is providing a structured buffer, Structured Constituent Buffer, so as to hold previous fillers for a case structure, instead of case registers before a verb appears in a sentence. The other is extended HOLD mechanism(in ATN), in whose use an embedded clause, especially a "be-deleted" clause, is recursively analyzed by case analysis. This parser's features are (1)extracting a case filler, basically as a noun phrase, by ATNG-based analysis, including recursive case analysis, and (2)mixing syntactic and semantic analysis by using case frames in case analysis.

1. Introduction

In a lot of natural language processing including machine translation, ATNG-based analysis is a usual method, while case analysis is commonly employed for Japanese language processing.The parser described in this paper consists of two major parts. One is ATNG-based analysis for getting case elements and the other is case-analysis for getting a semantic clause analysis. LUTE-EJ parser has been implemented on an experimental machine translation system LUTE (Language Understander, Translator & Editor) which can translate English into Japanese and vice versa. LUTE-EJ is the English-to-Japanese version of LUTE.

In case analysis, two ways are generally used for parsing. One way analyzes a sentence from left to right, by using case registers. Case fillers which fill each case registers are major participants of constituents, for example SUBJECT, OBJECT, PP(Prepositional Phrase)'s and so on, in a sentence. In particular, before a verb appears, at least one participant(the subject) will be registered, for example, in the AGENT register.

The other method has two phases on the analysis processing. In the first processing, phrases are extracted as case elements in order to fill the slots of a case frame. The second is to choose the adequate case element among the extracted phrases for a certain case slot and to continue this process for the other phrases and the other case slots. In this method, there are no special actions, i.e. no registering before a verb appears.(Winograd [83])

English question-answering system PLANES (Waltz [78]) uses a special kind of case frames, "concept case frames". By using them, phrases in a sentence, which are described by using particular "subnets" and semantic features (for a plane type and so on), are gathered and an action of a requirement (a sentence) is constructed.

2. LUTE-EJ Parser

2.1. LUTE-EJ Parser's Domain

The domain treated by LUTE-EJ parser is what might be called a set of "complex sentences and compound sentences". Let S be an element of this set and let CLAUSE be a simple sentence (which might include an embedded sentence). Now, if MAJOR-CL and MINOR-CL are principal clause and subordinate clause, respectively, S can be written as follows.

(R1) <S>::=(<MINOR-CL>)<MAJOR-CL>
(<MINOR-CL>)
(R2) <MAJOR-CL>::= <CLAUSE> / <S>
(R3) <MINOR-CL>::=<CONJUNCTION>
<CLAUSE> (in BNF)

The syntactic and semantic structure for a CLAUSE is basically expressed by a case structure. In this expression, the structure can be described by using case frames. The described structure implies the semantic structure intended by a CLAUSE and mainly depending on verb lexical information.

Case elements in a CLAUSE are Noun Phrases, object NPs of PPs or some kinds of ADVerbs with relation to times and locations. The NP structure is described as follows,

(R4)<NP>::=(<NHD>){<NP>/NOUN}(<NMP>)
/<Gerund-PH> /<To-infinitive-PH> /That<CLAUSE>

where NHD(Noun HeaDer) is "premodification" and NMP(Noun Modifier Phrase) is "postmodification". Thus, NMP is a set including various kinds of embedded finite clauses, relative or be-deleted relative finite clauses.

2.2. LUTE-EJ Parser Overview

After morphological analysis with looking up words for an input sentence in the dictionary, an input sentence analysis is begun from left to right. Thus, after a verb has been seen, it makes progress to analyze a CLAUSE by referring to the case frame corresponding to the verb, as each slot in the case frame is filled with an NP or an object of PP. A case slot consists of three elements: one semantic filler condition slot and two syntactic and semantic marker slots. Here, a preposition is directly used as a syntactic marker. Furthermore, four pseudo markers, "subject", "object", "indirect-object" and "complement", are used. As a semantic marker, a so-called deep case is used (now, 41 ready for this case system). Then, LUTE-EJ Parser extracts the semantic structure implied in a sentence (S or CLAUSE) as an event or state instance created from a case frame, which is a class or a prototype. An NP is parsed by the ATNG-based analysis in order to decide a case slot filler (now, 81 nodes on this ATNG).

Next, the reason why the case analysis and ATNG-based analysis are merged will be stated. It has two main points.

One point is about the depth of embedded structures. For example, the investigation on the degree of a CLAUSE complexity resulted in the necessity to handle a high degree of complexity with efficiency. The NMP structure is also more complex. In particular, embedded VPs or ADJPHs appear recursively. Therefore, a recursive process for analyzing NP is needed.

The other point is about the representation of grammatical structures. Grammar descriptions should be easy to read and write. Representations by using case frames make rules of any kind for NMP very simple, describing no NMP contents.

In order to deal with the above two points, combining the case analysis with ATNG-based analysis solves those problems. Verbal NMP(VTYPE-NMP)s are dealt with by recursive case-analyzing

2.3. Structured Constituent Buffer

As mentioned above, syntactic and semantic structures are basically derived from a sentence by analyzing a CLAUSE. Analysis control depends on the case frame, when the verb has been just appearing in a CLAUSE. However until seeing the verb, all of the phrases, which may be noun phrases with embedded clauses, PPs or ADVs before the verb, must be held in certain registers or buffers.

Here, a new buffer, STRuctured CONstituent Buffer(STRCONB), is introduced to hold these phrases. This buffer has surface constituents structure, and consists of specific slots. There are two slot types. One is a register to control English analysis and the other is a buffer to hold some mentioned-above constituents. The first type has two slots ; one is similar to a blackboard and registers the names of unfilled-slots. The other stacks the names of filled-slots in order of phrase appearance and is used for backtracking in the analysis. The second slot type involves several kinds of procedures. One of the main procedures, "getphrase", extracts some candidates for the slot filler from the left side of a CLAUSE. It fills the slot with these candidates. This procedure takes one argument, which is a constituent marker, "prepositional-phrase", "noun-phrase" and so on (in practice, using each abbreviation).

For example, when the following sentence is given, the evaluation for "(getphrase 'preph)"in LISP returns one symbol generated for the head prepositional phrase, "In the machine language", and determines the slot filler.

(s1) "In the machine language each basic machine operation is represented by the numerical code that invokes it in the computer, and....."

However, if the argument is "verb", this procedure only tells that the top word of unprocessed CLAUSE is a verb. At that moment, the process of filling with slots in STRCONB ends. Then case analysis starts.

2.4. CLAUSE Analysis

After seeing a verb in a CLAUSE, that is, filling the verb slot in the STRCONB, the case analysis starts. When the parser control moves on the case frame, the analyzer falls to work in order to fill the first case slot, which is generally one for the constituent SUBJECT and for the case AGENT or INSTRUMENT, etc. in the semantic structure. This first slot is special, because the filler has already been predicted in the slot for SUBJECT in STRCONB. Therfore, the predicted phrase is tested to determine whether or not it satisfies the semantic condition of the first case slot. If it is good, the slot is filled with it as a case instance. The parser control moves to the next case slot and a candidate phrase for it is extracted from the remainder of the input sentence by invoking the function "getphrase" with NP-

argument. This slot is usually OBJECT, or obligatory prepositional phrase name if the verb is intransitive. Furthermore, the control moves to the next case slot to fill it, if the case frame has more slots, all of which are obligatory case slots. They are described in a meaning slot (whose value is a meaning frame) in a case frame, while optional case slots are united in a special frame.

The process to fill the case slots is continuing until the end of the case frame. Then, more than one candidate for a case structure may be extracted. More than one for an NP extracted by "getphrase" gives many case structures, because of the difference in input remainders.

Next, recusive parsing will be mentioned. In analyzing embedded clauses, which are VTYPE-NMPs, CLAUSE analysis also gets in use of NPs parsing. It is supported with a new STRCONB. The procedure to call NP analysis is described in the next section. The conceptual diagram for LUTE-EJ analysis as a recusive CLAUSE is shown in Fig.1.

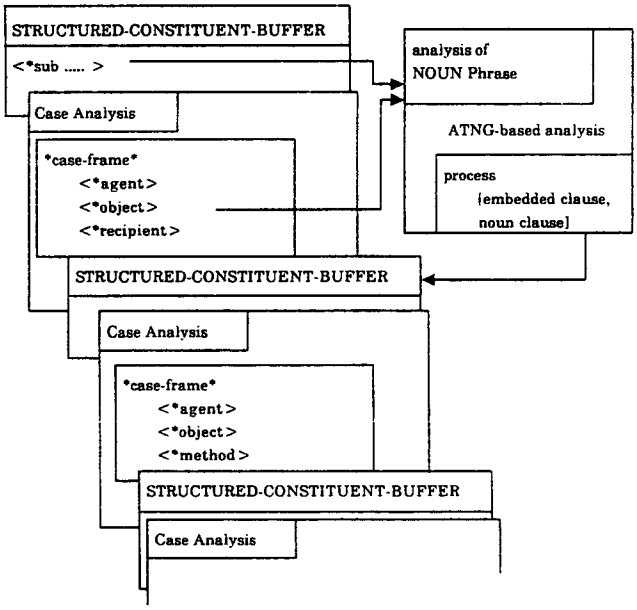

Fig.1 Conceptual Diagram of LUTE-EJ Analysis

2.5. NP Analysis

An NP structure is basically described as the rule (R4). In this paper, NHD structure and the analysis for it are omitted. NMP is another main NP constituent and will be explained here.

NMP is described in the following form.

(R5) <NMP> : : =
<PP> / <PResent-Participle-PHrase> /
<PaSt-Participle-PH> / <ADJective-PH> /
<INFinitive-PH> / <RELative-PH> /
<CARDINAL> <UNIT> <ADJ>

If an NMP is represented by any kind of VP or ADJ-PH, it is described in a case structure by using a case frame. That is, VTYPE-NMPs are parsed in the same way as CLAUSEs. However, a VTYPE-NMP has one (or more) structural missing element (a hole) compared with a CLAUSE. Therefore, complementing them is needed by restoring a reduced form to the complete CLAUSE. Extending "HOLD"-manipulation in ATN makes it possible. This extension deals with not only relative clauses but also VTYPE-NMPs. That is, the phrases with a "whiz-deletion" in Transformational Grammar can be treated. ADJ-PHs can also be treated. For example, the following phrase is discussed.

(s2) "I know an actor suitable for the part."
nmp

In the above case, the deletion of the words, "who is", results in the complete sentence being the above representation. The extending HOLD-manipulation holds the antecedent of a CLAUSE with a VTYPE-NMP. Calling the case analysis recursively, the VTYPE-NMP is parsed by it. Each VTYPE-NMP has a specific type, PRP-PH, PSP-PH, INF-PH or ADJ-PH. Each of them looks for an antecedent, as the object or the subject: so that each is treated according to the procedure to decide the role of the antecedent and the omitting grammatical relation. Therefore, it is necessary to introduce one "context" representing VTYPE-NMP. The present extension demands the context with the antecedent and calls the case analysis.

The following structured representation describes a NOUN, as stated above.

```
(NOUN
   (*TYPE ($value (instance)))
   (*CATEGORY ($value ("semantic-category")))
   (*SELF ($value ("entry-name")))
   (*POS ($value (noun)))
   (*MEANING ($value ("each-meaning-frame-list")))
   (*NUMBER ($value ("singular-or-plural")))
   (*MODIFIERS ($value ("NHD-or-NMP-instance-list")))
   (*MODIFYING ($value ("modificand")))
   (*APPOSITION($value ("appositional-phrase-instance")))
   (*PRE ($value ("prepositional-phrase-instance")))
   (*COORD ($value ("coordinate-phrase"))) )
```

Each word with prefix "*" describes a slot name such as a case frame has. However many slots are prepared for holding pointers to represent a syntactic structure of an NP. The value for VTYPE-NMPs *MODIFIERS is a pair of VTYPE-NMPs and an individual verbal symbol, for example, "(PRP-PH verb*1)".

Complementing NP's structure, an appositional structure is introduced. It is described in *APPOSITION-slot and treated in the same way as NMPs. Those phrases are discriminated from another NMP by a pair of a delimiter "," and a phrase terminal symbol, or, in particular, by proper nouns.

A Coordinate conjunction is another important structure for an NP. There are three kinds of coordinates in the present NP rule. The first is between NPs, the second is NHDs, and the third is NMPs. The NP representation with that conjunction is described by an individual coordinate structure. That is, the conjunction looks like a predicate with any NPs as parameters, for example, (and NP1 NP2 ..... NPi). Therfore, the coordinate structure has "*COORDINATE-OBJECTS" and "*OBJ-CAT" slot, each of which is filled with any instanciated NP/NHD/NMP symbol or any coordinate type, respectively.

Some linguistic heuristics are needed to parse NPs, along with extracting as few inadequate NP structures as possible. Several heuristics are introduced into LUTE-EJ parser. They are shown as follows.

(1) Heuristics for a compound NP

"Getphrase" function value for an NP is the list of candidates for an adequate NP structure. The function first extracts the longest NP candidate from an input. In this analysis, its end word is separated from the remainder of the input by some heuristics,

(a) The top word in the remainder is a personal pronoun.
(b) Its end word has a plural form.
(c) Its top is a determiner.

These heuristics prevent the value from having abundant non-semantical structures.

(2) Heuristics by using contexts

When NP analysis is called when filling a case slot, the case-marker's value for it is delivered to NP analysis. This value is called "syntactic local context". It is useful in rejecting pronouns, which are ungrammatically inflected, by testing the agreement with the syntactic local context and the subject or the object. Another context usage is shown below. Assume that a phrase containing a coordinate conjunction "and", for example, is in a context which is an object or a complement, and the word next to the conjunction is a pronoun. If the pronoun is a subjective case, the conjunction is determined to be one between CLAUSEs. To the contrary, the pronoun being a objective case determines the conjunction to connect an NP with it.

(3) Apposition

Many various kinds of appositions are used in texts. Most of them are shown by N. Sager [80]. The preceding appositional structures are used.

3. LUTE-EJ Parser Merits

3.1. A Merit of Using Case Analysis

In two sentences, each having different syntactic structures, there is a problem involved in identifying each case by extracting semantic relations between a predicate and arguments (NPs, or NPs having prepositional marks). LUTE-EJ case analysis has solved this problem by introducing a new case slot with three components (Section 2.2.). For case frames in LUTE-EJ analysis containing the slots, an analysis result has two features at the same time. One is a surface syntactic structure and the other is a semantic structure in two slots. Therefore, many case frames are prepared according to predicate meanings and case frames are prepared according to predicate meanings and syntactic sentence patterns, depending on one predicate (verb).

An analysis example is shown for the same semantic structure, according to which there are three different syntactic structures. These three sentences are as follow (from Marcus [80]).

(s3) "The judge presented the prize to the boy."
(s4) "The judge presented the boy with the prize."
(s5) "The judge presented the boy the prize."

Three individual structures are obtained for each sentence and their meaning equivalence for each slot is proved by matching the fillers of case-instances and by doing the same for case-names.

Incidentally, a sentence containing another meaning of "present" is as follows. It means "to show or to offer to the sight", for example, in a sentence,

(s6) "They presented the tickets at the gate."

In this case, the "present" frame must prepare the obligatory "at" case slot.

3.2. An Effect of Combining Case Analysis with ATNG-based Analysis

The next section shows one application of the LUTE-EJ parser, which is a machine translation system. So, taking the translated sample sentence in Section 4., effective points in parsing are shown in this section. The sample sentence is as follows.

(s7) "In the higher-level programming languages the instructions are complex statements, each equivalent to several machine-language instructions, and they refer to memory locations by names called variables."

One point is NMP analysis method by recursive calling for case frame analysis. In the example, two

NMP phrases are seen.

(a) The phrase which is an adjective phrase and modifies "each", appositive

to the preceding "statements",

(b) The phrase which is a past participle phrase and modifies "names".

These phrases are analyzed in the same case frame analysis, except for the phrase deletion types (depending on VTYPE-NMP) appearing in them. The deleted phrases are the subject part and the object part respectively. Judging from the point of a parsing mechanism, extended HOLD-manipulation transports the deleted phrases, "each" and "names", with the contexts to the case frame analysis.

The other point is to hold undecided case elements in STRCONB. The head PP and the subject in the sentences, for example, are buffering until seeing the main verb.

4. An Application to Machine Translation

One of the effective applications can be shown by considering the NMP analysis with embedded phrases. These NMPs are represented by instances of actions, i.e. individual case frames which may be having an unfilled case slot. Applying LUTE-EJ parser to an automatic machine translation system, there may be a little problem in lacking the case slots information. The reason is because the lacking information can be thought of as being indispensable for a semantic structure in one language, for example a target language Japanese, in spite of having them in another languages, for example a source language English. The problem is the difference in how to modify a head noun by an NMP or an embedded clause.

In Japanese, a NOUN is often modified by an embedded clause in the following pattern.

"<predicate's arguments>* <predicate> NOUN"

; * representing recursive applications

Therefore, in Japanese, an NMP phrase represented by a case frame corresponds to an embedded clause and the verb of the frame corresponds to the predicate.

A translation example is shown in Fig.2.

References

Marcus, Mitchell P., "A Theory of Syntactic Recognition for Natural Language", MIT Press, 1980.

Sager, Naomi, "Natural Language Information Processing", Addison-Wesley, 1981.

Waltz, David L., "An English Language Question-Answering System for a Language Relational Data Base", CACM Vol.21, 1978.

Winograd, Terry, "Language as a Cognitive Process", Vol.1, Addison-Wesley, 1983.

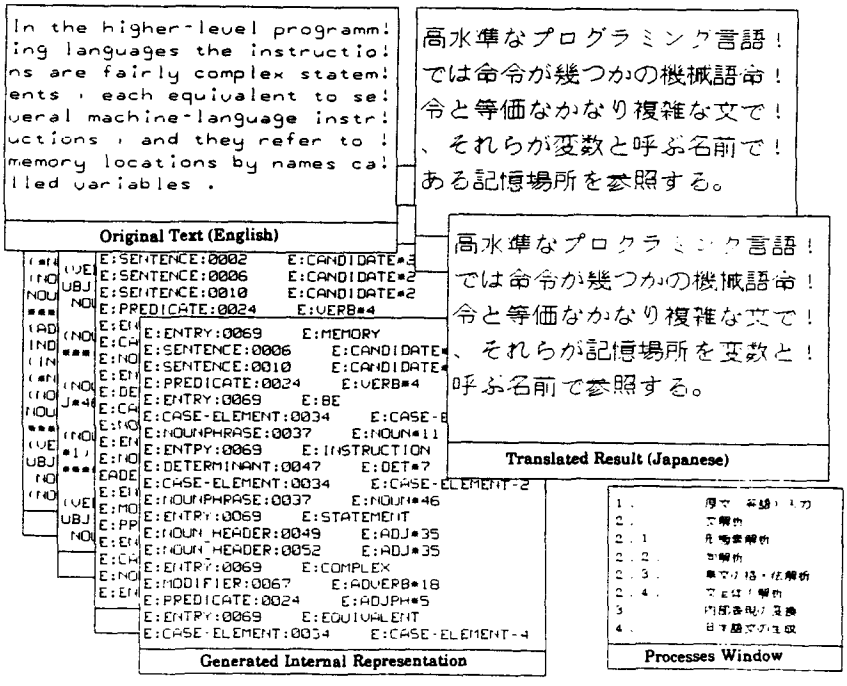

Fig. 2 An Example of LUTE Translation Results on the Display (from English to Japanese)