ICOT Technical Report: TR-603

TR-603

Disambiguation in natural language interpretation based on amount of information

by
K. Sumita, T. Ukita & S. Amano (Toshiba)

November, 1990

ICOT
Mita Kokusai Bldg. 21F
4-28 Mita 1-Chome
Minato-ku Tokyo 108 Japan
(03)3456-3191 ~ 5
Telex ICOT J32964

Institute for New Generation Computer Technology

Disambiguation in natural language interpretation based on amount of information*

Kazuo SUMITA, Teruhiko UKITA and Shin-ya AMANO
Toshiba Corp. R & D Center
Komukai-Toshiba-cho 1, Saiwaiku, Kawasaki 210, Japan

ABSTRACT

This paper describes how to use the amount of information in a sentence interpretation as a measure of interpreting input sentences in a natural language understanding system. In this paper, an interpretation of a sentence is considered to be a proposition, and the amount of information of the interpretation is defined according to a listener's model with a knowledge base composed of a literal set and a logical implication set, both of which are defined within the framework of propositional logic. When a given sentence can be analyzed syntactically and semantically into more than one interpretation, the most informative interpretation is selected. The theory of selecting the most informative interpretation by the proposed measure is reasonable in the sense that communication is an act whereby messages are passed on with the least possible effort. The presented theory for disambiguation is applied to a practical procedure for anaphoric ambiguity resolution, as an example of the disambiguation problem, which forms part of a question-answering system. Furthermore, a conversation experiment was carried out, and it was found that ninety-three percent of referents corresponding to anaphoric expressions could be correctly chosen.

1. Introduction

A number of ambiguities have to be confronted in understanding natural language by machine, including lexical ambiguity, structural ambiguity and contextual ambiguity. Lexical ambiguity refers to ambiguity of category, and ambiguity in identifying the concept of a word. For example, the concept *bank* is lexically ambiguous; the word *bank* in *Stay away from the bank* can mean either the land along a river or the place in which money is kept. Structural ambiguity refers to ambiguity with respect to the surface structure of a given sentence. *I saw a girl with a telescope* is a famous ambiguous sentence, which can be assigned two surface structures depending on whether *with a telescope* modifies *girl* or *saw*. Contextual ambiguity in a sentence arises from its relationship to the discourse context. A speaker and a hearer share context and knowledge, so that the speaker can use anaphora, such as pronouns, pro-verbs, definite noun phrases, and ellipses. Therefore the contextual ambiguity that anaphora causes can be reduced using the context and the knowledge. For example, the problem of determining the referent for *he* in *Fred phoned John because he needed help* is of this type. The word *he* can semantically refer to either *Fred* or *John*. According to these possibilities the statement *he needed help* therefore can be assigned two interpretations. The problem of resolving all of these ambiguity types can be viewed as that of a selection of the most plausible interpretation of an ambiguous sentence in a certain context. At the same time, this selection needs to be carried out using extra-linguistic knowledge as well as linguistic knowledge, because it is obvious that a human resolves these

* This work was supported by ICOT (Institute for New Generation Computer Technology), and was carried out as part of the fifth generation computer project in Japan.

ambiguities using such knowledge. This paper concerns disambiguation methods using extra-linguistic knowledge, and proposes an objective measure, based on such extra-linguistic knowledge, for the disambiguation of sentence interpretation in its discourse context. (In the following discussion, we simply use the term knowledge for this extra-linguistic knowledge.)

Several means for determining the most plausible interpretation in natural language understanding (NLU) have been investigated. Hobbs tried to solve the problem of reference, compound nominals and metonymy, by assessing the cost of assuming each condition of an inference rule to derive the conclusion[5]. Nishida[11] proposed an NLU mechanism using an assumption-based truth maintenance system (ATMS[2]). In this research, each possibility from the ambiguities corresponds to an assumption, and an assumability probability, which is defined as a heuristic measure, is used to control these assumptions. The accuracy of both theories depends on the intuitive magnitude of the cost or probability, but no formal definitions or computation methods for these measures are given. Thus these measures cannot be adopted as a general criterion for the disambiguation. A theory based on an objective measure for disambiguation needs to be developed.

To define the measure for disambiguation, a listener's comprehension model is constructed as a part of communication process between a speaker and a hearer. The comprehension model is based on the presumption that the speaker might transmit maximum information with minimum effort. The hearer can comprehend speaker's messages according to this presumption. If an amount of information in a sentence interpretation is defined, a machine hearer can employ this measure for resolving ambiguities in the comprehension process. We consider the presumption, which the comprehension model is founded on, to be relevant in natural conversation. For example Grice's conversational maxim of quantity also states that a speaker should make his contribution as informative as is required[3].

The amount of information in a knowledge representation, which is an expansion of *many sorted logic*, was discussed by Ohsuga[12]. The amount of information associated with a structure of objects for arguments of a predicate was defined in his research. This value seems to be a criterion appropriate for our use, because the NLU process can be viewed as a process of knowledge acquisition through comprehending each sentence in its discourse context. However, a means of practically calculating the amount of information including inference rules was not considered in his research, while the role of causal relationships between events represented by inference rules is important in an NLU process[16] . Moreover, disambiguating sentence interpretations in NLU needs a quantitative measure of information on a sentence interpretation or a statement. Thus it is necessary to consider a definition of amount of information in a statement against the background of a knowledge base with inference rules.

This paper defines the amount of information for interpreting a sentence based on a listener's comprehension model which includes knowledge about causal relationships between event, and adopts it as the criterion for disambiguation. In the following sections, a comprehension model and a theoretical framework for the amount of information in a sentence interpretation are described using propositional logic. This theory of amount of information is applied to anaphora resolution, which is a typical disambiguation problem. Finally, a conversation experiment is carried out

using a question answering system with a knowledge processing sub-system.

2. Listener's comprehension model

Figure 1 shows a listener's comprehension model including knowledge*1; the model consists of two major processes: sentence analysis and preference judgement. The first process chooses interpretation candidates corresponding to a new input sentence. In the second process, the most plausible interpretation is determined according to the knowledge base, and this interpretation is added to the knowledge base and treated as new knowledge.

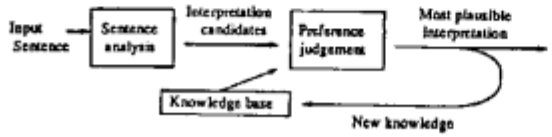

Fig.1 Listener's comprehension model.

In general, the knowledge base includes information of the so-called *world*, knowledge about the objective field, information of the speaker, etc. In this paper, we will deal with knowledge about the objective field, i.e., knowledge about facts and causal relationships in the field.

First, let's consider the listener's knowledge in terms of propositional logic. In the framework of propositional logic, all propositions are constructed from atomic propositions and logical connectives. Negation($\neg$), conjunction($\wedge$), disjunction($\vee$) and logical implication($\rightarrow$) form the logical connectives. An atomic proposition and the negation of an atomic proposition form a literal, and a proposition which is created by combining more than two literals is called a compound proposition. Note that a literal is not defined as a compound proposition in this paper. We assume that the listener's knowledge comprises a set of literals and a set of logical implications, i.e., inference rules. The set of literals corresponds to knowledge about facts and events with which the listener is acquainted through the discourse, and the set of logical implications corresponds to knowledge about known causal relationships. Let's define a set of all atomic propositions to provide a formal definition of the listener's knowledge. Only these atomic propositions can be used to form propositions in the listener's knowledge.

Definition 2.1. A set of all atomic propositions Ω is defined as follows,

$$\Omega=\{\omega_n | 1\leq n\leq N\}. \quad (1)$$ □

All propositions in listener's knowledge are constructed from atomic propositions in Ω using logical connectives, and are classified as a set of literals and a set of logical implications. It should be noted that all the atomic propositions ω_n in Ω need to be used to construct the set of propositions in the listener's knowledge.

Definition 2.2. The listener's knowledge HK consists of D and K (HK = D $\cup$ K). Here, D is the set of literals, where any element π_i in D is either $\pi_i=\omega_n$ or $\pi_i=\overline{\omega_n}$ for an element ω_n in Ω. And K is the set of logical implications, where any atomic proposition used in logical implications κ_j in K belongs to Ω. □

Each proposition in HK is consistent; an *interpretation* (an assignment of truth value to all atomic propositions) for the proposition exists. In other words, each proposition in HK holds true for the listener who knows HK, and therefore the truth value of the conjunction of propositions in HK is true for him.

It can be said, without loss of generality, that every logical implication in K is an

*1 A listener means a question answering system, and a speaker means a human user of the system.

implication such that the left-hand side is a conjunction of literals, and the right-hand side is a single literal, because the following theorem is provable.

Theorem 2.1. For any D and K, the sets D and K can be translated into logically equivalent sets D1 and K1. D1 is a set of literals, and K1 is a set of logical implications, where the left-hand side of any logical implication in K1 is a conjunction of literals, and the right-hand side is a single literal.
(The proofs of all theorems are given in the Appendix.) □

In the subsequent discussion, we assume that every logical implication in K is in a form such that the left-hand side is a conjunction of literals, and the right-hand side is a single literal. The contents of D and K can be written as follows,

$$D = \{\pi_i \mid 1 \le i \le M\}, \quad (2)$$
$$K = \{\kappa_j \mid 1 \le j \le J\}. \quad (3)$$

Next, let's describe the comprehension process based on the knowledge defined above. A sentence including ambiguous factors can be broken down into several possible interpretations. In this paper, the interpretation of a sentence is dealt with as a proposition. Here, let's denote the possible interpretation as ν. When ν is added to the listener's knowledge, some new propositions are derivable from ν and the listener's knowledge (D and K), and a new literal set D' is constructed.

Definition 2.3. Let ν be a new input proposition. After ν becomes known to the listener, the new set of literals comes to be known to him. This set is denoted by D'. Then D' is defined as follows,

$$D' = D \cup \{\pi'_i \mid (\pi'_i \in \Omega \text{ or } \overline{\pi'_i} \in \Omega) \text{ and } \nu, D, K \vdash \pi'_i\}, \quad (4)$$

where $\nu, D, K \vdash \pi'_i$ means the literal π'_i can be derived from ν, D, K. □

Example 2.1. Suppose that D and K are given as follows,

$$D = \{\pi_1, \pi_2, \pi_3\} = \{\omega_1, \omega_4, \overline{\omega_6}\}.$$
$$K = \{\kappa_1, \kappa_2\} = \{\omega_3 \wedge \omega_4 \rightarrow \omega_5, \ \omega_1 \wedge \overline{\omega_2} \rightarrow \omega_3\}.$$

When $\nu = \overline{\omega_2}$ is a new proposition corresponding to an input sentence, the new proposition set D' will be $\{\omega_1, \overline{\omega_2}, \omega_3, \omega_4, \overline{\omega_6}\}$. □

3. Amount of information in an interpretation

This section gives the formal definition of amount of information in a sentence interpretation on the basis of the comprehension model within the framework described in section 2.

3.1 Definition of amount of information in an interpretation

When a proposition corresponding to an input sentence is added to the listener's knowledge, the proposition alters the possible states of the listener's knowledge. Figure 2 shows the states of the listener's knowledge before and after he hears a statement ("The button was pushed"). Before he heard the statement, he did not know whether the button was pushed or not; therefore there were two even possibilities of the situation in his knowledge state. On the other hand, after he had heard the statement, the possibility that the button was not pushed was discarded; only one possibility remains. Therefore, the function of an input sentence can be described as the reduction of the possible states of a listener's knowledge.

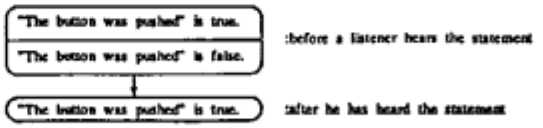

Fig.2 States of listener's knowledge before and after he has heard a statement.

To formalize this process, let's consider the possible combinations of all atomic propositions with truth values. The possible combinations are denoted by the sequence of all atomic propositions in Ω with truth values, which are enclosed in < and >. For instance, if all atomic propositions other than ω_1 are true, and ω_1 is false, then the combination can be written by $< \overline{\omega_1}, \omega_2, \omega_3, \ldots, \omega_N >$. The state of the listener's knowledge is defined as follows:

Definition 3.1. The state of the listener's knowledge can be expressed by the set of all possible combinations of an atomic propositions with truth values consistent with D and K*2. The state of the listener's knowledge is denoted by $S_\Omega(D,K)$. (In the subsequent discussion, the subscript Ω in $S_\Omega(D,K)$ will be omitted to simplify the description, because Ω is invariant throughout the discussion of amount of information.) □

Example 3.1. Let's consider a case where $D=\phi$ and $K=\phi$. Since all possible combinations regarding affirmation or negation in terms of all atomic propositions are consistent with D and K, the state $S(\phi,\phi)$ exists as follows.

$$\begin{aligned} S(\phi,\phi)=\{&< \omega_1, \omega_2, \omega_3, \ldots, \omega_N >,\\ &< \overline{\omega_1}, \omega_2, \omega_3, \ldots, \omega_N >,\\ &< \omega_1, \overline{\omega_2}, \omega_3, \ldots, \omega_N >,\\ &\ldots\\ &< \overline{\omega_1}, \overline{\omega_2}, \overline{\omega_3}, \ldots, \overline{\omega_N} >\}. \end{aligned}$$ □

*2 As for a given proposition, when the conjunction of all elements in the literal set D and the proposition is true under a certain assignment of truth values to all atomic propositions, we say that the proposition is consistent with D. Also, when the conjunction of all elements in the inference set K and the proposition is true under a certain assignment of truth values to all atomic propositions, we say that the proposition is consistent with K.

Here, we regard the truth value of a conjunction of literals contained within an element in S(D,K) as that of the element. In general, each element in S(D,K) is mutually exclusive, i.e., if a certain element is true, then the other elements cannot be true. At the same time, it is not possible to identify a specific element in S(D,K) as true, i.e., it is not known which element holds true. Therefore, the state S(D,K) can be viewed as a sample space in information theory[13]. If a sample space can be partitioned in a finite number of mutually exclusive elements, whose probabilities p_i are assumed to be known, then the measure of uncertainty or *entropy* associated with the space is given by

$$-\sum_{i=1}^{I} p_i \log p_i,$$

where I is the number of elements of the space. In general, because each element in the state S(D,K) can be expected to occur in equal probability, we make the following assumption.

Assumption 3.1. The prior probability of every possible state in S(D,K) is constant, and all are equal to each other.

$$p_i = |S(D,K)|^{-1}, \text{ for all } i,$$

where |S(D,K)| means the number of elements in S(D,K). □

Using Assumption 3.1, the entropy of S(D,K) can be obtained, and the entropy of the listener's knowledge can be defined as the entropy of S(D,K).

Definition 3.2. The entropy of the listener's knowledge, E(D,K), is defined as the entropy of S(D,K).

$$\begin{aligned} E(D,K) &= -\sum_{n=1}^{|S(D,K)|} |S(D,K)|^{-1} \log_2 |S(D,K)|^{-1}\\ &= \log_2 |S(D,K)|, \qquad (5) \end{aligned}$$

When a new proposition ν is given, and when D is replaced by D' obtained from (4), the amount of information in the new proposition ν is provided by I(ν,D,K).

$$I(\nu,D,K)=E(D,K) - E(D',K) = \log_2 (|S(D,K)| / |S(D',K)|). \quad (6)$$ □

This equation indicates that the amount of information in a proposition is calculated by the number of elements in the state of the listener's knowledge before and after the proposition is added.

In the case of knowledge without inference rules, Ohsuga discussed a similar derivation for the amount of information in a proposition by quantitative consideration of knowledge representation[12]. Here, the derivation of amount of information as described in his work is described briefly according to our notation.

Example 3.2. Because his theory does not contain inference rules, $K=\phi$ in our notation. For the case $D=\phi$, it is clear that $E(\phi,\phi)=N$, since $|S(\phi,\phi)|=2^N$. For instance, suppose that a new proposition added to the proposition set D ($=\phi$) is ω_1. Then, $D'=\{\omega_1\}$. Because $S(\{\omega_1\},\phi)$ consists of only one possible combination of all atomic propositions with truth values consistent with ω_1, clearly $|S(\{\omega_1\},\phi)|$ is half of $|S(\phi,\phi)|$. Thus, $|S(\{\omega_1\},\phi)|=2^{N-1}$, and $I(\{\omega_1\},\phi))=1$. Also, as to other propositions ν such that $\nu=\omega_n$ or $\nu=\overline{\omega_n}$ for $\omega_n\in\Omega$, it is clear that if $\nu\notin D$ and $K=\phi$, then $A=\phi$ and $D'=D\cup\{\nu\}$. Therefore $|S(D',\phi)|$ is half of $|S(D,\phi)|$. As a result, the amount of information in a proposition ν, $I(\nu,D,\phi)$, is equal to 1. □

3.2 Computation of amount of information in the case of knowledge with inference rules

(a) General case

Let's discuss the case of knowledge with inference rules. In general, some inference rules are dependent on each other; i.e., we can suppose that there are sets of inference rules which share common propositions. First, inference rules are classified into equivalence classes, each of which consists of inference rules sharing common atomic propositions. Second, all combinations of atomic propositions with truth values in each equivalence class are considered to constitute the entropy of the listener's knowledge.

For the purpose of the classification of inference rules, a relation R for rules in K and an equivalence relation R' are introduced.

Notation 3.1. For each κ_i and κ_j in K, $\kappa_i R \kappa_j$, if and only if an atomic proposition ω_n exists, which both κ_i and κ_j contain as components. □

Then let's define the equivalence relation R' based on R.

Notation 3.2. The equivalence relation on elements in K which satisfies the following two conditions is denoted by R'.

(i) For all i and j, if $\kappa_i R \kappa_j$ then $\kappa_i R' \kappa_j$

(ii) For all i, j and k, if $\kappa_i R' \kappa_j \wedge \kappa_j R' \kappa_k$ then $\kappa_i R' \kappa_k$ □

An equivalence relation on a set can divide the set into several sets which do not have any element in common. Thus, using the relation R', Lemma 3.1 is stated.

Lemma 3.1. Equivalence classes Γh ($\bigcup_h \Gamma h = K$, $\Gamma h \cap \Gamma k = \phi$ for $h \neq k$) exist, and each element in K can be assigned as a member of one of these Γh ($1\le h\le H$, where H is the total number of the classes) by the relation R'. □

Γh is the set of inference rules which are connected directly or indirectly with each other, and the intersection of Γi and Γj is empty for any i and j ($i \neq j$). Therefore, when a combination of atomic propositions

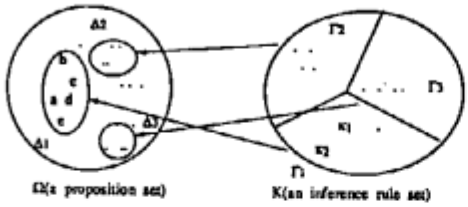

Fig.3 Relation between an inference rule set Γh and a proposition set Δh (κ1 is a∧b→c, and κ2 is c∧$\bar{d}$→e.)

√a,b,c,d,e	√a,$\bar{b}$,c,d,e	√$\bar{a}$,b,c,d,e	√$\bar{a}$,$\bar{b}$,c,d,e
a,b,c,d,$\bar{e}$	a,$\bar{b}$,c,d,$\bar{e}$	$\bar{a}$,b,c,d,$\bar{e}$	$\bar{a}$,$\bar{b}$,c,d,$\bar{e}$
√a,b,c,$\bar{d}$,e	√a,$\bar{b}$,c,$\bar{d}$,e	√$\bar{a}$,b,c,$\bar{d}$,e	√$\bar{a}$,$\bar{b}$,c,$\bar{d}$,e
×a,b,c,$\bar{d}$,$\bar{e}$	×a,$\bar{b}$,c,$\bar{d}$,$\bar{e}$	×$\bar{a}$,b,c,$\bar{d}$,$\bar{e}$	×$\bar{a}$,$\bar{b}$,c,$\bar{d}$,$\bar{e}$
×a,b,$\bar{c}$,d,e	√a,$\bar{b}$,$\bar{c}$,d,e	√$\bar{a}$,b,$\bar{c}$,d,e	√$\bar{a}$,$\bar{b}$,$\bar{c}$,d,e
×a,b,$\bar{c}$,d,$\bar{e}$	a,$\bar{b}$,$\bar{c}$,d,$\bar{e}$	$\bar{a}$,b,$\bar{c}$,d,$\bar{e}$	$\bar{a}$,$\bar{b}$,$\bar{c}$,d,$\bar{e}$
×a,b,$\bar{c}$,$\bar{d}$,e	√a,$\bar{b}$,$\bar{c}$,$\bar{d}$,e	√$\bar{a}$,b,$\bar{c}$,$\bar{d}$,e	√$\bar{a}$,$\bar{b}$,$\bar{c}$,$\bar{d}$,e
×a,b,$\bar{c}$,$\bar{d}$,$\bar{e}$	a,$\bar{b}$,$\bar{c}$,$\bar{d}$,$\bar{e}$	$\bar{a}$,b,$\bar{c}$,$\bar{d}$,$\bar{e}$	$\bar{a}$,$\bar{b}$,$\bar{c}$,$\bar{d}$,$\bar{e}$

Fig.4 Combinations of affirmations and negations for all atomic propositions (a,b,c,d,e) (Combinations with × are not consistent with either a∧b→c or c∧$\bar{d}$→e. Combinations with √ are not consistent with $\bar{e}$.)

is considered, relationships between inference rules belonging to other classes do not have to be taken into account. Corresponding to Γh, atomic propositions Δh can be defined as subsets of Ω; Δh is the set of atomic propositions which appear in inference rules in Γh. Figure 3 shows this relation between Γh and Δh. Let's consider all combinations of atomic propositions in Δh with truth values in a similar manner as the notion of S(D,K) for K,

Notation 3.3. Let C(D,Γh) be a set of all combinations of atomic propositions in Δh with truth values, which are consistent with D and Γh, where Δh is a set of atomic propositions which appear in Γh. □

Thus, Theorem 3.1 is proved.

Theorem 3.1. If a proposition v relates to an inference rule*3 in Γh, the amount of information in v can be calculated by combinations of all atomic propositions in Δh with truth values, regardless of the other classes.

$$I(v,D,K) = \log_2 (|C(D,\Gamma h)| / |C(D',\Gamma h)|), \quad (v \text{ is related to } \kappa j \text{ in } \Gamma h \text{ and } v \notin D), \tag{7}$$

where D', given by (4), is the set of literals derived from v, D and K. □

The amount of information in any proposition v can be calculated in accordance with the above theorem, and is given by

$$I(v,D,K) = \begin{cases} \log_2 (|C(D,\Gamma h)| / |C(D',\Gamma h)|) \\ \quad (v \text{ is related to } \kappa j \text{ in } \Gamma h \text{ and } v \notin D), \\ 1 \quad (v \text{ is not related to any } \kappa j \\ \qquad \text{in K and } v \notin D), \\ 0 \quad (v \in D). \end{cases} \tag{8}$$

To illustrate the calculation process, let's consider an example:

Example 3.3. Consider the following case:

$\Omega = \{a, b, c, d, e, f, g, h, p, q, r\}$,
$D = \{f\}$,
$K = \{\kappa_1, \kappa_2, \kappa_3, \kappa_4\}$
$= \{a \wedge b \rightarrow c,\ c \wedge \bar{d} \rightarrow e,\ f \wedge g \rightarrow h,\ g \rightarrow q\}$,
$v = \bar{e}$.

In this example, because κ_1 and κ_2 have c, and κ_3 and κ_4 have g in common, respectively, K can be divided into Γ1 and Γ2. Then Γ1, Γ2, Δ1 and Δ2 are given as follows,

$\Gamma 1 = \{\kappa_1, \kappa_2\}$,
$\Gamma 2 = \{\kappa_3, \kappa_4\}$,
$\Delta 1 = \{a, b, c, d, e\}$,
$\Delta 2 = \{f, g, h, q\}$,
(Propositions "r" and "p" do not belong to either Δ1 or Δ2).

Since the new proposition, the negation of "$\bar{e}$", is equal to the right-hand side of κ_2, "$\bar{e}$" relates to κ_2 in Γ1. Thus, the combinations with truth values in terms of propositions in Δ1 corresponding to Γ1 should be taken into account, regardless of Δ2 and Γ2. The matrix depicted in Figure 4 shows

*3 When either v or $\bar{v}$ exists in the set of atomic propositions which are the components of some inference rule κ_j in K, we say that the proposition v is related to the inference rule κ_j.

these combinations; there are 32 combinations in the matrix.

As a first step, combinations not consistent with either κ_1 or κ_2 are removed, and these are marked with '×" in Figure 4, after which 24 combinations remain. In this case, because no elements in D are contained in Δ_1, the 24 combinations are consistent with D.

Secondly, combinations not consistent with the proposition $v=\bar{e}$ are removed from these 24 combinations, and they are marked with "√ " in Figure 4.

The terms now remaining without marks form combinations consistent with Γ_1 and D', and it is easy to see that $|C(D',\Gamma_1)|=10$. In consequence, the amount of information in "e" is obtained; $I(e,D,K)=\log_2(24/10)=1.26$.

(b) Special case (where individual inference rules do not share common propositions)

In the general case, it is required to consider all true combinations of literals which are directly or indirectly linked to an input proposition. Therefore, to calculate the amount of information in the proposition a large amount of memory or CPU time is needed. For a practical calculation procedure, let's consider a special case where no other inference rules have atomic propositions in common. If the inference rule set K satisfies the following condition, then it is not necessary to consider the truth value combinations of all the atomic propositions.

Assumption 3.2. For each κ_i and κ_j in K, where $i\neq j$, there is no common proposition which is shared by both κ_i and κ_j. □

On the basis of Assumption 3.2, the amount of information in a proposition v depends only on the inference rule κ_j which v relates to, and $|C(D,\Gamma_h)|$ in Theorem 3.1 can be simplified, where $\Gamma_j = \{\kappa_j\}$ by the subscript h of Γ_h corresponding to the subscript j of κ_j.

Theorem 3.2. If v is related to κ_j in Γ_j, then $|C(D,\Gamma_j)|$ on the basis of Assumption 3.2 is given as follows,

$$|C(D,\Gamma_j)|=2^{m_j}-k_j-\sigma_j(D), \qquad (9)$$

where m_j is the number of literals in κ_j, k_j is the number of literals (in κ_j) whose truth values are determined in D, and $\sigma_j(D)$ is a function which gives 1 when the negation of κ_j is consistent with D, otherwise 0. □

As a result, (8) is simplified as follows,

Theorem 3.3. If v is related to κ_j, then the amount of information in v on the basis of Assumption 3.2 is given as follows,

$$I(v,D,K)=\begin{cases}\log_2\{(2^{m_j}-k_j-\sigma_j(D)) / (2^{m_j}-k_j-1-\sigma_j(D'))\} \\ \quad (v \text{ is related to some } \kappa_j \text{ and } v\notin D), \\ 1 \quad (v \text{ is not related to any } \kappa_j \text{ and } v\notin D), \\ 0 \quad (v\in D). \end{cases} \qquad (10)$$ □

Example 3.5. Consider the following case:

$D = \{a\}$,
$K = \{a\wedge b\wedge c\rightarrow d,\ e\wedge f\rightarrow g\}$,
$v = b$.

Since $\kappa_j \equiv a\wedge b\wedge c\rightarrow d$, the negation of κ_j is $\kappa_j = a\wedge b\wedge c\wedge\bar{d}$. Each parameter can be calculated; $m_j=4$, $k_j=1$. Both $\sigma_j(D)$ and $\sigma_j(D')$ are 1, since κ_j is consistent with D and D'. In consequence, $I(v,D,K)=\log_2(7/3)=1.22$. □

In this section, the amount of information in a proposition has been defined on the basis of knowledge including the inference rules; its magnitude depends on the possible combinations of truth values of atomic propositions. In a special case where a proposition is linked to proposi-

tions by an inference rule defined in Assumption 3.2, the amount of information in the proposition is calculated by a simple operation on the few proposition sets which the added proposition relates to.

We can apply this theory on amount of information to disambiguate the interpretation of a given natural language sentence, because the proposition corresponds to one of the possible interpretations of a sentence accepted by the NLU system. Therefore, amount of information as defined by (8) can be viewed as a means of assigning an amount of information to a sentence interpretation.

Equation (8) states that if an interpretation is linked with more facts or events which have been introduced during the discourse, and if more guesses are derived from the known facts and events, then $|C(D',\Gamma h)|$ decreases. As a result, the amount of information is larger. Thus, we can conclude that the selection of a interpretation maximizing the amount of information corresponds to the use of inference rules in knowledge for ambiguity resolution.

However, the use of the amount of information given by (8) as a measure for disambiguating sentence interpretations in a practical NLU system may not be advantageous, because large amounts of memory and CPU time are required to calculate it using (8). For a practical calculation procedure, we have considered a special case in which individual inference rules do not share common propositions. Equation (10) gives a good enough approximation of the amount of information in a given proposition in the general case, unless inference rules share many common propositions. In the following section, we apply this equation to disambiguating interpretations in an NLU system.

4. Application to natural language understanding

We will describe an application of the theory in disambiguating interpretations in a practical NLU system. In section 3, it was shown that the amount of information in a sentence interpretation based on a comprehension model including knowledge can be used to resolve ambiguity in its discourse context. By selecting an interpretation with the largest amount of information, the interpretation having the strongest link with the discourse context can be determined.

We will deal with the anaphora resolution problem as a typical example of a disambiguation problem, because this problem is basic and important in natural language understanding, and because an anaphoric expression used in some discourse contexts gives rise to contextual ambiguity. Anaphora is an intersentential device which connects a superficial expression with an object or event in the discourse, such as by the use of pronouns, pro-verbs, certain definite noun phrases, or ellipses. The problem of anaphora resolution is that of determining the referent of an anaphoric expression. When more than two referent candidates satisfying the semantic restrictions of an anaphoric expression exist, it is necessary to decide which candidate is the most plausible. The formalization for amount of information described in the previous sections is applied to the problem of resolving anaphoric ambiguity by regarding the problem of anaphora resolution as one of sentence interpretation.

4.1 Problem

Figure 5 shows an example sentence in a consultation dialogue on the operation of a video cassette recorder (VCR)[15]. Consider the anaphoric expression "it" in this example. The referent candidate "VCR" is preferable in a consultation on

VCR operation to other candidates "cassette-tape" or "playback-button", though all of these candidates satisfy the semantic restriction of "work", too. This preference comes from the fact that the listener can be supposed to have the following knowledge, if he has used a VCR.

If a VCR is on,
and someone inserts a cassette tape into the VCR,
and pushes the playback button,
then the VCR will work.

This causal relationship embodied in the listener's knowledge is needed in order to select the most plausible candidate, because this ambiguity cannot be solved by using syntactic and/or semantic restrictions alone. The theory presented in the previous sections supports the use of this causal relationship: selecting the informative interpretation based on the listener's knowledge leads to its use.

Though I inserted a tape into the VCR
and pushed the playback button, it's not working.

Fig.5 An example of a sentence in a dialogue about VCR operation.

4.2 Procedure

To manage the problem described above, a practical procedure is presented; the procedure consists of three major steps: anaphora detection; referent candidate extraction; and preference judgement. In the preference judgement step the procedure chooses the appropriate referent for an anaphora in the input sentence through selecting the most informative interpretation of the sentence.

Every input sentence in Japanese, which is written in *kanji* and *kana*, is transformed into a dependence structure of case-frames representing an interpretation. After the input sentence is analyzed morphologically into a sequence of *bunsetsus*[18], the case-frames are constructed from the dependencies of nearest *bunsetsu* pairs that satisfy semantic restrictions, where any cross-serial dependency is inhibited. For each case-frame, the following procedure is carried out.

Anaphora detection

As the first step in anaphora resolution, the procedure detects anaphoric expressions, such as pronouns, definite nouns, omitted obligatory cases of the predicate in the case-frame being processed, and nouns themselves*4. These extracted expressions are managed as tentative instance schemata. □

Referent candidate extraction

For every anaphoric expression, the procedure is to search for referent candidates corresponding to the expression. Instance schemata which belong to the same class as the expression, those which belong to a sub-class, and those which belong to a class linked to the expression's by a has-part relationship, are extracted as candidates. On the other hand, if an anaphoric expression refers to some event, the procedure searches for preceding events in the conversation history. After extracting referent candidates, the procedure constructs a structural representation for the case-frame being processed by binding the referent candidates to the verb of the case-frame. This representation includes a list of interpretation candidates for the sentence, and in general there is a number of these candidates. □

Preference judgement

When there is more than one interpretation candidates in the preceding processes, the following procedure is followed. The procedure is to choose the most informative interpretation; preceding sentences

*4 A noun in a Japanese sentence can refer to an object or an event in the discourse.

in the conversation history are inspected in order to link the interpretation candidate being processed with some preceding sentences by if-then rules in the system's knowledge base. Here the group of if-then rules corresponds to the set of inference rules K, and the group of preceding sentences in conversation history corresponds to the set of literals D in our theory. An interpretation candidate which maximizes the amount of information calculated by Equation (10) is selected as the most plausible one. □

After the procedure selects the interpretation candidates for all case-frames, it accepts the next sentence. Figure 6 shows a process example for the sentence: "VCR no saisei botan wo oshita ga ugokanai", (in English it means "though I push the VCR playback button, it's not working"). The second clause is analyzed into interpretation candidates v_1 and v_2, and the calculated amount of information of

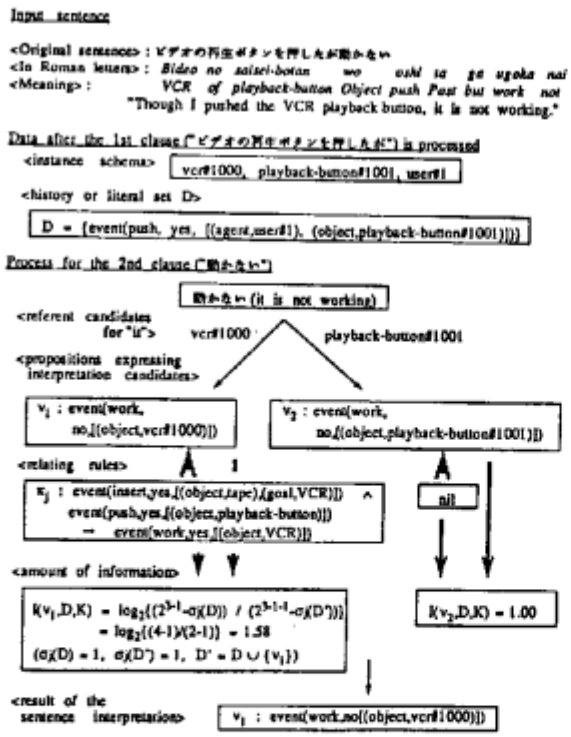

Fig.6 An example of the procedure for disambiguation of anaphoric references.

v_1 and v_2 is 1.58 and 1.00, respectively. Therefore the proposition v_1 is selected as the preferable interpretation for that clause, and at the same time, the referent for *it* is determined to be the VCR.

The procedure has been incorporated as a part of an experimental question-answering system, whose current task is guidance in VCR operation[15]. The system uses a knowledge representation system[9], which represents an object and an event as a *schema* in a similar way as *units* in KRL[1] and causal relationships between events by if-then rules. This Q/A system has been developed on PSI-II*5, and has 1000 words in its dictionary, and 100 event schemata, 200 object schemata, and 30 rules in its knowledge base. It can be said that the number of elements in the inference rule set K of our theory is 30, because the above if-then rules correspond to the elements in K. On the other hand, the number of propositions constructed from event schemata and object schemata cannot be counted because of the large number of combinations, but this is no problem, since we do not have to pay attention to the number of atomic propositions in set Ω to calculate the amount of information in a given proposition.

4.3 An Experiment

A conversation experiment has been carried out using the experimental Q/A system with the procedure for anaphora resolution described above. This experiment was aimed at an evaluation of the theory presented here through the anaphora resolution procedure, not at an evaluation of the total Q/A system. Here, the following restricted conversation experiment

*5 PSI-II is a computer system developed by the Institute for New Generation Computer Technology.

was carried out. Nine pairs of sentences, such as those shown in Figure 7, were given to four persons, and sentences which could follow these pairs were collected from them.

User: ビデオの再生方法を教えろ。
"Show me how to playback a tape."
System: ビデオカセットテープをビデオに入れてから再生ボタンを押して下さい。
"Please insert the tape into the VCR, and push the playback button."

(a) An example of given dialogues.

ビデオに入れる前にボタンを押すとどうなるか。
"If I push the button before I insert it, what will happen?"

(b) An example of a collected sentence which can follow (a).

Fig.7 An example of given dialogues and collected sentences.

Several of these sentences couldn't be processed correctly because of the use of non-registered words, so any sentences that the system was unable to analyze syntactically were removed; the number of sentences remaining was 113, and they were used in this experiment (there were 174 clauses in these sentences). After given dialogues, these sentences were input into the system. There were 204 anaphoric expressions in the collected sentences, and 190 referents were found correctly. For example the following dialogue was processed correctly.

System

Bideo-kasetto-teipu wo *bideo ni* *ire masita ka.*
video-cassette-tape Object VCR Goal insert Past ?
Did you insert the Video cassette tape into the VCR?

User

[ϕ*wo*] *Ire* *ta* *ga,* [ϕ*ga*] *dete kite simau.*
insert Past but [it] came out
I inserted it, but it came out of the VCR. □

The video-cassette-tape is plausible for the ellipsis ϕ*ga* in [ϕ*ga*] *dete kite simau*, though ϕ*ga* could refer to both the picture on the screen and the VCR within a syntactic and semantic analysis.

The procedure produced more than one interpretation candidate 47 times, and it failed to select the correct candidate 6 of those times. These errors were due to inadequacy in the candidate extraction step. For example, the procedure failed to find correct objects, because they were too far from anaphoric expressions to be found as candidates; the number of objects which the procedure searched for was limited to 10 preceding objects, because of the processing time. On the other hand, all interpretations related to an if-then rule were selected correctly, and this occurred 32 times. As a result, 91 percent of interpretations were chosen correctly in total (93 percent of referents were determined correctly).

4.4 Related Work on Anaphora Resolution

Several works on anaphora resolution from various points of view have been presented (e.g., Ref. (14) and (17)). They fall into three broad categories:

- approaches using general heuristics to find referents,
- approaches using syntactic and semantic restrictions to eliminate
- approaches using the inference method to infer cognitive entities or to verify coherency with the context.

Our approach can be categorized as an the inference approach, however it uses semantic restrictions, too.

Many studies demonstrate that a theory for anaphora resolution must accommodate the role not only of syntactic and semantic restrictions but also of inferential knowledge. In one of these studies, a forward chaining method of inferences including a "demon mechanism" provides entities as referents for anaphoric expressions, where the entities are not lexical elements in the text[10]. The role of inferences, however, is not limited to this function; inferences are needed to solve anaphoric ambiguity. Hobbs tried to solve anaphoric ambiguity by using coherence relations between sentence pairs[6]; in his

study, anaphoric expressions are interpreted as having the function of binding the sentence pairs. The relationships are categorized into contrasts, parallels, and elaborations, and they are recognized by using inference rules.

There are also several studies using some kind of focusing mechanisms (e.g., Ref. (4), (7), (14) and (8)). In a typical approach, Sidner proposed a bootstrapping procedure using a focusing mechanism; a focus serves as a primary referent as long as it leads the context to a coherent one. In her approach, the inference function simply verifies the consistency of the focused objects within the context.

These works using some inference functions do not present any objective quantitative measure for selecting the most plausible referent. They cannot answer the question of why a particular inference rule is objectively appropriate for linking a sentence with its context. The formal theory presented in this paper gives a theoretical background for the use of inference functions in anaphora resolution. The sentence interpretation linked to its context by an inference rule contains more information than other interpretations not linked to its context.

5. Conclusion

The amount of information in a sentence interpretation, used as a measure for disambiguation in a natural language understanding system, is described. It is defined on the basis of a listener's comprehension model including knowledge about causality relations between events as well as propositions. A theory for calculating the amount of information in a possible interpretation is described. Furthermore, this measure is applied to a procedure for anaphora disambiguation, which forms part of a question-answering system. Finally, a conversation experiment was carried out using this system. Ninety-three percent of referents were determined correctly. Even though many sentences could not be used in the experiment, the effectiveness of the defined measure was confirmed; the error rate on interpretation decisions was reduced by around half, and is expected to be reduced still further.

This paper presents the first work based on formal definitions of the listener's knowledge and gives a quantitative measure for resolving ambiguities in natural language, especially anaphoric ambiguities. There are several works using inference methods for dealing with anaphoras, and the work of Hobbs used an inference method to resolve some anaphoric ambiguities[6]. These works are similar to ours from the viewpoint of their use of inference methods; however they do not provide any formal quantitative definition in terms of the use of this inferential knowledge; they simply used inference methods. Our work presents a theoretical background for the use of inference methods in anaphora resolution.

The theory has been developed within the framework of propositional logic. Propositional logic is adequate to disambiguate interpretations in such simple Q/A system as do not deal with quantifiers; and the result of the experiment proves this. Propositional logic is, however, too simple to deal with natural language completely, because it neglects linguistic elements such as the mood, aspect, or tense of a verb. Furthermore, the structure of the real knowledge of a listener might not be flat unlike the set of propositions defined in this paper. In further study, we intend to expand the framework in these areas.

Appendix.

Theorem 2.1.

Proof. To prove this theorem, we use the notion that the conjunction of

propositions in HK (= D ∪ K) is true. If any logical implication exists in K, it can be represented by a conjunction, whose conjuncts are literals or disjunctions of literals, because any compound proposition can be transformed to a conjunctive normal form. Thus, the conjunction of propositions in HK is transformed into a conjunctive normal form. Each conjuncts of the conjunctive form can be separated into other elements in K1 or D1 respectively, i.e., literals belong to D1, and disjunctions belong to K1. Furthermore any disjunction can be rewritten as a logically equivalent logical implication, and it is clear that the left-hand side of the logical implication is a conjunction of literals, and the right-hand side is a single literal. □

Lemma 3.1.

Proof. It is clear that the relation R' over K is an equivalence relation, because R' is reflective, symmetric and transitive from Notation 3.1 and 3.2. Therefore, K is grouped into equivalence classes by the equivalence relation R'. □

Theorem 3.1.

Proof. No two proposition sets Δi and Δj for $i \neq j$, share any common proposition by definition. Also, let the number of conjunctive combinations of those which don't belong to any Δi be C0 ($C0=2^{N0}-L$, where N0 is the number of propositions which don't belong to any Δi, and L is the number of atomic propositions, other than those propositions whose truth values are determined in D already.) Then, the number of elements in S(D,K) is obtained as the product of the elements in combinations $C(D,\Gamma h)$;

$$|S(D,K)| = C0 \times \prod_{i=1}^{H} |C(D,\Gamma i)|. \qquad (A.1)$$

Only $C(D,\Gamma h)$ with respect to Γh which the input proposition v relates to is variant, before and after v is added to D. Thus, substituting (A.1) into (6) gives (7). □

Theorem 3.2.

Proof. Let Ψj and $\Psi' j$ be sets of literals corresponding to a inference rule κ_j:

$$\Psi j=\{\overline{\rho_{j1}},\overline{\rho_{j2}},\ldots,\overline{\rho_{j(kj-1)}},\rho_{jkj}\}, \qquad (A.2)$$

$$\Psi' j=\{\rho_{j1},\rho_{j2},\ldots,\rho_{j(kj-1)},\overline{\rho_{jkj}}\}, \qquad (A.3)$$

where $\kappa_j \equiv \rho_{j1} \wedge \rho_{j2} \wedge \ldots \wedge \rho_{j(kj-1)} \rightarrow \rho_{jkj}$. $D \cap (\Psi j \cup \Psi' j)$ is the set of all literals in D, and also related to κ_j, because $\Psi j \cup \Psi' j$ is the set of all literals in terms of atomic propositions in κ_j. Thus, $|D \cap (\Psi j \cup \Psi' j)| (=kj)$ is the number of literals (in κ_j) whose truth values are determined in D. The number of all combinations of atomic propositions with truth values in κ_j is 2^{mj}, therefore the number of combinations consistent with D is $2^{mj} - kj$. Of these combinations, only the combination $<\rho_{j1},\rho_{j2},\ldots,\rho_{j(kj-1)},\overline{\rho_{jkj}}>$ is not consistent with κ_j, so the total number of the combinations is reduced by one, when $D \cap (\Psi j \cup \Psi' j)$ contains the possibility of $<\rho_{j1},\rho_{j2},\ldots,\rho_{j(kj-1)},\overline{\rho_{jkj}}>$, i.e., when $D \cap (\Psi j \cup \Psi' j) \subset \Psi' j$. $D \cap (\Psi j \cup \Psi' j) \subset \Psi' j$ means that the negation of κ_j is consistent with D. Therefore, $|C(D,\Gamma j)|=2^{mj-kj-1}$ when the negation of κ_j is consistent with D, otherwise $|C(D,\Gamma j)|=2^{mj-kj}$. □

Theorem 3.3.

Proof. By substituting (9) into (8), (10) is obtained. □

REFERENCES

(1) D. Bobrow and T. Winograd: "KRL: Knowledge Representation Language", Cognitive Sci., Vol.1, No.1 (1977).

(2) J. de Kleer: "An Assumption-based Tms", Artif. Intell., Vol.28, pp.127-162 (1986)

(3) H..P. Grice: "Logic and Conversation", Syntax

and Semantics, Vol.3, Speech Acts, Academic Press, pp.41-58 (1975).

(4) B.J. Grosz: "The Representation and Use of Focus in a System for Understanding Dialogs", Proc. 5th IJCAI '77, pp.67-76 (1977).

(5) J.R. Hobbs, M. Stickel, P. Martin and D. Edwards: "Interpretation as Abduction", Proc. 26th ACL '88, pp.95-103 (1988).

(6) J.R. Hobbs: "Coherence and Coreference", Cognitive Sci., Vol.3, pp.67-90 (1979).

(7) A.K. Joshi and S. Weinstein: "Control of Inference of Some Aspects of Discourse Structure - Centering", Proc. 7th IJCAI '81, pp.385-387 (1981).

(8) M. Kameyama: "A Property-sharing Constraint in Centering", Proc. 24th ACL '86, pp.200-206 (1986).

(9) S. Kinoshita, H. Sano, T. Ukita, K. Sumita and S. Amano: "Knowledge Representation and Reasoning for Discourse Understanding", Logic Programming '88, K.Furukawa, et al. (Eds.), Springer-Verlag, pp.238-251 (1989).

(10) M. Nagao, J. Tsujii and K. Tanaka: "Analysis of Japanese Sentences, by Using Semantic and Contextual Information - Context Analysis", Trans. IPS Japan, Vol.17, No.1, pp.19-28 (1976) (in Japanese).

(11) T. Nishida, X. Liu, S. Doshita, A. Yamada: "Maintaining Consistency and Plausibility in Integrated Natural Language Understanding", Proc. 12th COLING '88, pp.482-487 (1988).

(12) S. Ohsuga: "A Consideration to Knowledge Representation - an Information Theoretic View", Trans. IPS Japan, Vol.25, No.4, pp.685-694 (1984) (in Japanese).

(13) F.M. Reza: "An Introduction to Information Theory", McGraw-Hill Book Company (1961).

(14) C.L. Sidner: "Focusing in Comprehension of Definite Anaphora", Computational Models of Discourse, M.Brady, et al. (Eds.), MIT press, pp.267-330 (1983).

(15) T. Ukita, K. Sumita, S. Kinoshita, H. Sano, S. Amano: "Preference Judgement in Comprehending Conversational Sentences using Multi-paradigm World Knowledge", Proc. Int. Conf. Fifth Generation Computer Systems '88, pp.1133-1140 (1988).

(16) T. Ukita and S. Kinoshita: "Knowledge Representation and Reasoning for Natural Language Understanding - Knowledge Representation", J. IPS Japan, Vol.30, No.10, pp.1224-1231 (1989) (in Japanese).

(17) B.L. Webber: "Syntax beyond the Sentence: Anaphora", Theoretical Issues in Reading Comprehension, R.J.Spiro, et al. (Eds.), Lawrence Erlbaum Associates, Inc. (1980).

(18) H. Sano, R. Sugimura, K. Akasaka and Y. Kubo: "A Morphological Analysis based on Word Formation", IPS Japan Technical Report, NL 66-3 (1988) (in Japanese).